Biotechnology in Agriculture and Forestry

Edited by T. Nagata
H. Lörz and J.M. Widholm

57 Plant Metabolomics

Edited by K. Saito, R. A. Dixon, and L. Willmitzer

Springer

Biotechnology in Agriculture and Forestry

Edited by
T. Nagata (Managing Editor)
H. Lörz
J. M. Widholm

Biotechnology in Agriculture and Forestry

Volumes already published and in preparation are listed at the end of this book.

Biotechnology in Agriculture and Forestry 57

Plant Metabolomics

Edited by
K. Saito, R.A. Dixon, and L. Willmitzer

With 96 Figures, 29 in Color, and 10 Tables

Springer

Library of Congress Control Number: 2005936763

ISSN 0934-943X
ISBN-10 3-540-29781-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29781-9 Springer Berlin Heidelberg New York

Springer is a part of Springer Science + Business Media
springer.com

Printed in Germany

Editor: Dr. Dieter Czeschlik, Heidelberg, Germany
Desk Editor: Dr. Andrea Schlitzberger, Heidelberg, Germany
Cover design: *design&production* GmbH, Heidelberg, Germany
Typesetting and production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig, Germany
Printed on acid-free paper 31/3152 5 4 3 2 1 0

Preface

Metabolomics is a rapidly-emerging sector of post-genome research. The metabolome (a set of all metabolites of an organism) represents not only the ultimate phenotype of cells by the perturbation of gene expression and the modulation of protein functions caused by the environment or mutations, but the metabolome can also feed back on gene expression and protein function. Therefore, metabolomics plays a key role for understanding cellular systems. Metabolomics is applied to a variety of biological fields from medical science to agriculture. Nevertheless, metabolomics research is particularly important in the plant field because plants collectively produce a huge variety of chemical compounds, far more than animals and even microorganisms. The number of all metabolites in the plant kingdom is estimated at 200,000 or more. In addition, most of the human-beneficial properties of plants, be they foods, medicinal resources, or industrial raw materials, are ascribed to plant metabolites.

This book aims to review the current status of plant metabolomics research. Since metabolomics itself is a new field, no such comprehensive book has yet been published. The chapters are divided into three sections: analytical technology, bioinformatics, and applications. These represent three major elements of metabolomics research. Each chapter provides cutting-edge information contributed by leading researchers from throughout the world.

We hope that this book will be a landmark for plant metabolomics research into the future and will give beneficial guidance to graduate students and researchers in academia, industry, and technology transfer organizations. Since metabolomics is still a growing discipline, further technology development in chemical analysis and bioinformatics will be required. We look forward to breakthrough technology innovations in metabolomics, and yet unforeseen findings and applications in plant science.

Finally, we would like to acknowledge our contributors who have enthusiastically put their efforts to ensure the high scientific quality of this volume. We also would like to thank our colleagues at Springer.

January 2006

Kazuki Saito,
Richard A. Dixon,
and Lothar Willmitzer

Contents

Section II Bioinformatics

List of Contributors

L. Achnine
Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

P. Ahiahonu
Phenomenome Discoveries Inc., 204–407 Downey Road, Saskatoon, Saskatchewan, Canada S7N 4L8

M. Altaf-Ul-Amin
Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Takayama-cho 8916–5, Ikoma, Nara 630–0101, Japan

M. Arita
Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, 277–8561 Japan, e-mail: arita@k.u-tokyo.ac.jp

H. Asahi
Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Takayama-cho 8916–5, Ikoma, Nara 630–0101, Japan

M.H. Beale
The National Centre for Plant and Microbial Metabolomics, Rothamsted Research, West Common, Harpenden, Herts. AL5 2JQ, UK, e-mail: mike.beale@bbsrc.ac.uk

R.J. Bino
Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

C. Böttcher
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany

P.M. Bramley
School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK

F. Carrari
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476, Golm, Germany

Y.H. Choi
Division of Pharmacognosy, Section Metabolomics, Institute of Biology, Leiden University, P.O. Box 9502, 2300RA, Leiden, The Netherlands

S. Clemens
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany, e-mail: sclemens@ipb-halle.de

R.B. Croteau
Institute of Biological Chemistry, Washington State University, Pullman, WA 99164, USA

T. Daskalchuk
Phenomenome Discoveries Inc., 204–407 Downey Road, Saskatoon, Saskatchewan, Canada S7N 4L8, e-mail: info@phenomenome.com

B.E. Deavours
Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

R.A. Dixon
Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA, e-mail: radixon@noble.org

A.R. Fernie
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476, Golm, Germany, e-mail: fernie@mpimp-golm.mpg.de

H. Foerster
Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA

M. Franz
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany

P.D. Fraser
School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK, e-mail: p.bramley@rhul.ac.uk

Y. Fujiwara
Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, 277–8561 Japan

E. Fukusaki
Department of Biotechnology, Graduate School of Engineering, Osaka Univ, 2–1 Yamadaoka, Suita, 565–0871, Japan,
e-mail: fukusaki@bio.eng.osaka-u.ac.jp

I.A. Graham
CNAP, Department of Biology (Area 7), University of York, PO Box 373, York YO10 5YW, UK

J. Gullberg
Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden

R.D. Hall
Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands, e-mail: robert.hall@wur.nl

D. Heath
Phenomenome Discoveries Inc., 204–407 Downey Road, Saskatoon, Saskatchewan, Canada S7N 4L8

M.Y. Hirai
RIKEN Plant Science Center, Suehiro-cho 1–7–22, Tsurumi-ku, Yokohama, Kanagawa 230–0045, Japan

T. Hirayama
International Graduate School of Arts and Sciences, Yokohama City University, 1-7-29 Suehiro, Tsurumi-ku, Yokohama, 230-0045 Japan

T. Ikegami
Department of Polymer Science and Engineering, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto, 606–8585, Japan

A.I. Johansson
Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden

P. Jonsson
Research Group for Chemometrics; Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

K.-M. Oksman-Caldentey
VTT Biotechnology, P.O. Box 1500, 02044 VTT, Finland,
e-mail: Kirsi-Marja.Oksman@vtt.fi

S. Kanaya
Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Takayama-cho 8916–5, Ikoma, Nara 630–0101, Japan,
e-mail: skanaya@gtc.naist.jp

R.E.B. Ketchum
Institute of Biological Chemistry, Washington State University, Pullman, WA 99164, USA, e-mail: rketchum@wsu.edu

J. Kikuchi
RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045 Japan, e-mail: kikuchi@psc.riken.jp

H.K. Kim
Division of Pharmacognosy, Section Metabolomics, Institute of Biology, Leiden University, P.O. Box 9502, 2300RA, Leiden, The Netherlands,
e-mail: verpoort@chem.leidenuniv.nl

J. Kopka
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany, e-mail: Kopka@mpimp-golm.mpg.de

K. Kurokawa
Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Takayama-cho 8916–5, Ikoma, Nara 630–0101, Japan

T.R. Larson
CNAP, Department of Biology (Area 7), University of York, PO Box 373, York YO10 5YW, UK, e-mail: trl1@york.ac.uk

B. Mehrotra
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington St., MC 0477, Blacksburg, Virginia 24061, USA

P. Mendes
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington St., MC 0477, Blacksburg, Virginia 24061, USA, e-mail: mendes@vt.edu

T. Moritz
Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden, e-mail: thomas.moritz@genfys.slu.se

Y. Nakamura
Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Takayama-cho 8916–5, Ikoma, Nara 630–0101, Japan

Y. Nakanishi
Intec Web and Genome Informatics Corporation

M. Naoumkina
Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

D. Ohta
Department of Plant Genes and Physiology, Graduate School of Agriculture and Biological Sciences, Osaka Prefecture University, Gakuen-cho 1–1, Sakai, Osaka 599–8531, Japan

M. Orešič
VTT Biotechnology, P.O. Box 1500, 02044 VTT, Finland

S.Y. Rhee
Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA, e-mail: rhee@acoma.stanford.edu

C.H. Ric de Vos
Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

H. Rischer
VTT Biotechnology, P.O. Box 1500, 02044 VTT, Finland

K. Saito
Chiba University, Graduate School of Pharmaceutical Sciences, Yayoi-cho 1-33, Chiba 263-8522, Japan
RIKEN Plant Science Center, Suehiro-cho 1–7–22, Tsurumi-ku, Yokohama, Kanagawa 230–0045, Japan, e-mail: ksaito@faculty.chiba-u.jp

N. Sakurai
Kazusa DNA Research Institute, 2–6–7 Kazusa-kamatari, Kisarazu, Chiba 292–0818, Japan

N. Schauer
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476, Golm, Germany

D. Scheel
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany

D. Shibata
Kazusa DNA Research Institute, 2–6–7 Kazusa-kamatari, Kisarazu, Chiba 292–0818, Japan, e-mail: shibata@kazusa.or.jp

Y. Shinbo
New Energy and Industrial Technology Development Organization, Toshima, Tokyo 170–6028, Japan

L.W. Sumner
The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA, e-mail: lwsumner@noble.org

H. Suzuki
Kazusa DNA Research Institute, 2–6–7 Kazusa-kamatari, Kisarazu, Chiba 292–0818, Japan

N. Tanaka
Department of Polymer Science and Engineering, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto, 606–8585, Japan

C. Tissier
Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA

T. Tohge
RIKEN Plant Science Center, Suehiro-cho 1–7–22, Tsurumi-ku, Yokohama, Kanagawa 230–0045, Japan

T. Tokimatsu
Kazusa DNA Research Institute, 2–6–7 Kazusa-kamatari, Kisarazu, Chiba 292–0818, Japan

R.N. Trethewey
metanomics GmbH and metanomics Health GmbH, Tegeler Weg 33, 10589 Berlin, Germany, e-mail: richard.trethewey@metanomics.de

J. Trygg
Research Group for Chemometrics; Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

E. v. Roepenack-Lahaye
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany

H.A. Verhoeven
Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

R. Verpoorte
Division of Pharmacognosy, Section Metabolomics, Institute of Biology, Leiden University, P.O. Box 9502, 2300RA, Leiden, The Netherlands

J.L. Ward
The National Centre for Plant and Microbial Metabolomics, Rothamsted Research, West Common, Harpenden, Herts. AL5 2JQ, UK,
e-mail: Jane.ward@bbsrc.ac.uk

L. Willmitzer
Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476, Golm, Germany

E. Willscher
Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany

Y. Yamazaki
Phenomenome Discoveries Inc., 204–407 Downey Road, Saskatoon, Saskatchewan, Canada S7N 4L8

P. Zhang
Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA

I.1 Gas Chromatography Mass Spectrometry

J. Kopka[1]

1 Introduction

GC-MS technology has been used for decades in studies which aim at the exact quantification of metabolite pool size and metabolite flux. Exact quantification has traditionally been focused on a single or small set of predefined target metabolites. Today GC-MS is one of the most widely applied technology platforms in modern metabolomic studies. Since early applications in unravelling the mode of action of herbicides (Sauter et al. 1988) it has experienced a renaissance (Fig. 1) in post-genomic, high-throughput fingerprinting and metabolite profiling of genetically modified (e. g. Roessner et al. 2001a,b, 2002; Fernie et al. 2004) or experimentally challenged plant samples (e. g. Cook et al. 2004; Kaplan et al. 2004; Urbanczyk-Wochniak and Fernie 2005). Metabolic phenotyping and analysis of respective phenocopies by metabolite profiling has become an integral part of plant functional genomics (Fiehn et al. 2000b; Roessner et al. 2002; Fernie et al. 2004). The essence of metabolite profiling, namely the non-biased screening of biological samples for changes of metabolite levels relative to control samples, has been thoroughly discussed earlier and is clearly distinguished from fingerprinting approaches and the concept of exact quantification (Fiehn et al. 2000b; Sumner et al. 2003; Birkemeyer et al. 2005).

GC-MS-based metabolome profiling analysis is on the verge of becoming a routine technology. This fact substantially contributes to the development of metabolomics as a fourth integral part of the Rosetta stone for functional genomics and molecular physiology (Trethewey et al. 1999; Fiehn et al. 2000b; Trethewey 2004). Nevertheless, GC-MS technology is already challenged again by new bottlenecks and demands for improved data sets which are optimised for the mathematical modelling tools currently developed in the fields of bioinformatics and biological systems analysis.

The challenges of modern, multi-parallel, GC-MS based metabolite analysis are manifold: (i) automation of sample preparation, wet chemistry and data processing after acquisition for increased throughput and reproducibility, (ii) extension of the analytical scope of metabolomics studies, for example by combined analysis of single samples using multiple analytical technology platforms, and combined analysis with the proteome and transcriptome

[1] Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany, e-mail: Kopka@mpimp-golm.mpg.de

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

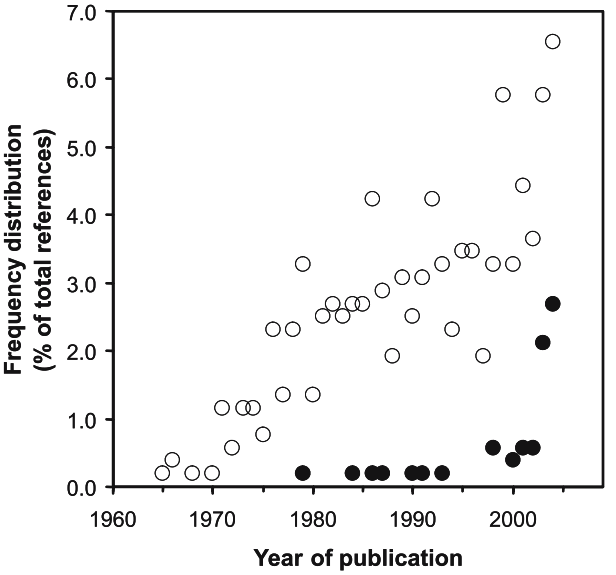


Fig. 1. Literature survey of publications which associate the concepts, "metabolite", "profiling", and "gas chromatography" performed on 1/2005. A total of ~500 citations without conference proceedings, abstracts and book chapters were found. The frequency of publications in all biological sciences (*open circles*) is compared to the contribution by plant metabolomics community (*closed circle*)

(Weckwerth et al. 2004b), (iii) profiling of trace compounds, or signalling molecules in the presence of bulk metabolites (Mueller et al. 2002; Birkemeyer et al. 2003; Schmelz et al. 2003, 2004), (iv) increasing accuracy in multi-parallel metabolite quantification (Birkemeyer et al. 2005), (v) combining profiling and flux analyses (Roessner-Tunali et al. 2004), (vi) establishment of quantitative repeatability, unambiguous nomenclature and comparability between analyses performed in different laboratories or using different analytical technology platforms (Schauer et al. 2005), and (vii) finally – perhaps the most important challenge of all metabolomic investigations – the identification of the unidentified majority of metabolic components from metabolite profiling experiments (Fiehn et al. 2000a; Schauer et al. 2005).

In agreement with the focus of this chapter the above challenges have predominantly analytic or technical motivation. The breakthrough of metabolomic investigations, however, will depend on the access to hitherto unavailable fundamental insights into metabolic and systems interactions. Increasingly integrative studies which consider the metabolome, proteome, transcriptome, and genome evolution of an organism have been initiated and are to be expected. Promising steps have been made – using GC-MS technology – towards network analysis (Fiehn 2003; Weckwerth et al. 2004a) and correlation studies between or within metabolome and transcriptome

constituents (Urbanczyk-Wochniak et al. 2003; Steinhauser et al. 2004; Kopka et al. 2005). A detailed discussion of these general aspects including GC-MS studies and beyond can be found in the applications section of this book.

2 GC-MS Profiling Technology in a Nutshell

Metabolite profiling with GC-MS involves six general steps:

1. *Extraction* of metabolites from the biological sample, which should be as comprehensive as possible, and at the same time avoid degradation or modification of metabolites (e. g. Kopka et al. 2004).
2. *Derivatisation* of metabolites making them amenable to gas chromatography. Metabolites which are not volatile per se require chemical modification prior to GC analysis.
3. *Separation* by GC. High resolution GC can also be highly reproducible as it involves automated sample injection robotics, highly standardised conditions of gas-flow, temperature programming, and standardised capillary column material.
4. *Ionisation* of compounds as they are eluted from the GC. Electron impact (EI) ionisation is most widely used, as it is the technology which is least susceptible to suppression effects and produces reproducible fragmentation patterns.
5. Time resolved *detection* of molecular and fragment ions. Mass separation and detection can be achieved with different mass-detection devices, including sector field detectors, quadrupole detectors (QUAD), ion trap technology, and time-of-flight detectors (TOF). The choice of detectors depends on the targeted analytical niche. GC-MS systems with QUAD detection are most widely spread for routine analysis. Ion trap technology allows MS×MS (two-dimensional MS) analysis for structural elucidation and targeted quantification of trace compounds (e. g. Mueller et al. 2002). TOF detection can either be tuned to fast scanning rates (van Deursen et al. 2000) or to high mass precision comparable to sector field systems. Fast scanning GC-TOF-MS enables the, today, most advanced technology in the GC-MS field, namely two dimensional GC×GC-TOF-MS (two-dimensional GC-TOF-MS) (Ryan et al. 2004; Sinha et al. 2004a–c).
6. *Acquisition and evaluation* of GC-MS data files. All GC-MS system manufacturers provide software which is tuned for targeted, quantitative metabolite analysis. The targeted approach involves unequivocal identification of predefined metabolites by expected chromatographic retention times and mass-spectral fragmentation patterns and quantitative calibration by authentic standard concentrations. Recent software developments support the non-targeted analysis of GC-MS patterns, and the full evaluation of all resolved compounds. This feature of GC-MS allows discovery of novel hitherto

unknown metabolites. As we are far from knowing all possible metabolites of a given organism, non-biased, truly comprehensive data evaluation is the most essential requirement of metabolite profiling.

2.1 Chemical Derivatisation and Chromatography

The principles of fast metabolic sample inactivation and nondestructive extraction are common to all metabolome analyses. In contrast to all other technologies GC-MS is inherently restricted to volatile and temperature-stable compounds. The scope of GC-MS for metabolite analysis is limited by the typical temperature range of commercial capillary columns, for example up to 320–350 °C. The lower temperature range is determined by ambient temperature, but cold trapping devices and isothermal GC allow analysis of low molecular weight gases and highly volatile metabolites. GC received a considerable extension of applications through the development of a highly versatile tool box of derivatisation reagents, which chemically transform non-volatile metabolites into volatile analytes for GC-MS analysis (e. g. Knapp 1979; Blau and Halket 1993; Toyo'oka 1999). To date, GC-MS profiling of metabolites in plants has largely been confined to compounds, recovered in the methanol-water phase after methanol-water/chloroform extraction of tissues (Fiehn et al. 2000a; Roessner et al. 2000; Duran et al. 2003; Barsch et al. 2004; Gullberg et al. 2004; Strelkov et al. 2004; Broeckling et al. 2005). Although not all hydrophilic compounds can be volatilised by derivatisation, the following classes of compounds are detected routinely: amino-, organic-, and aromatic-acids, amines, sugars up to trisaccharides, alcohols and polyols, and some mono-phosphorylated metabolites.

The current limitations of metabolite preparation and derivatisation strategy, namely methoxyamination with subsequent direct trimethylsilylation of predominantly polar metabolites, call for extension. Application of other technology platforms is an obvious route and will be discussed in the following chapters. Here a short appraisal of the potential of chemical derivatisation is attempted. Four main types of reaction schemes will be discussed.

1. *Alkoxyamination* by reagents, such as methoxyamine CH_3-O-NH_2, stabilises carbonyl moieties in native metabolite structures, but forms E- and Z-isomers of the $-N=C<$ double-bond substituents. Keto-enol tautomerism is suppressed, as is the decarboxylation of unstable β-carbonyl-carboxylic acids. In addition, the formation of acetal- or ketal-structures in aqueous solution is inhibited. These equilibrium reactions generate multiple intramolecular and water adducts, for example the typical α- and β-conformers of reducing sugars. Ether- and ester-conjugates are mostly stable when exposed to methoxyamine reagent and maintain conformation. So far other alkoxy-reagents – for example hydroxylamine, ethyloxyamine, or benzyloxyamine – have not been exploited for systematic discovery of metabolites with carbonyl moieties:

2. *Silylation* reagents classify into those which introduce either a trimethylsilyl (TMS) moiety, $-Si(CH_3)_3$, or a dimethyl-(*tert*-butyl)-silyl (TBS) moiety, $-Si(CH_3)_2-C(CH_3)_3$. TMS reagents have been well investigated and are known to have the widest derivatisation spectrum (Little 1999; Halket et al. 2005). TMS has the potential to substitute all exchangeable, "acidic" protons of a metabolite. Steric hindrance of TMS substitution is rare but common with the bulkier TBS reagent. The benefit of the TBS reagent is higher tolerance for the presence of water and clear mass spectral fragmentation. However, vicinal diols, which typically occur in sugars, are only partially derivatised.
3. *Alkylation* reactions, mostly methylation, are widely used to derivatise carboxylic acids and alcohols. The enormous reactivity of available reagents – some allow for flash derivatisation during hot GC injection – leads to transalkylation of ester-bonds und consequently breaks down complex metabolites, such as glycero- and phospholipids. Alkylation of sugars leads to derivatives which are more volatile than the TMS derivatives and therefore allow analysis of higher sugar oligomers.
4. *Acylation* reactions, mostly acetylation or trifluoro-acetylation, are less reactive than transalkylation. Reagents usually form stable ester and amide bonds and break down only activated metabolic intermediates, e. g. thioesters.

In conclusion further developments of alternate GC-MS profiling techniques need to employ more selective combinations of metabolite fractionation and derivatisation schemes. Solid phase extraction can be explored to partition and concentrate metabolites amenable to alternate subsequent derivatisation. On the other hand, vapour phase extraction (VPE) for the separation and concentration of volatile derivatisation products prior to GC injection may prove promising (Schmelz et al. 2003, 2004). VPE has the potential to be a robust technique and was shown to operate with a range of commonly used reagents.

2.2 Mass Detection and Quantitative Calibration Techniques

One of the major criticisms and pitfalls of metabolome analyses is best explained by so-called matrix effects. This well-known effect describes unexpected losses or increased recovery of metabolites in complex extracts compared to pure authentic preparations. Matrix effects on one hand are caused by the presence of compounds which either specifically inhibit extraction or chemical analysis of metabolites. Positive matrix effects can stabilise otherwise labile compounds in the presence of suitable chemicals. Typical examples are suppression effects of soft ionization techniques, for example electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI). Electron-impact ionization (EI) typically used in GC-MS profiling is not susceptible to suppression. Instead GC injection is the crucial step which may

cause discriminations, especially in view of the complex and rather crude extracts which are typically injected.

So far, only exemplary – albeit time demanding – thorough tests for unexpected matrix effects have been performed with selections of chemically

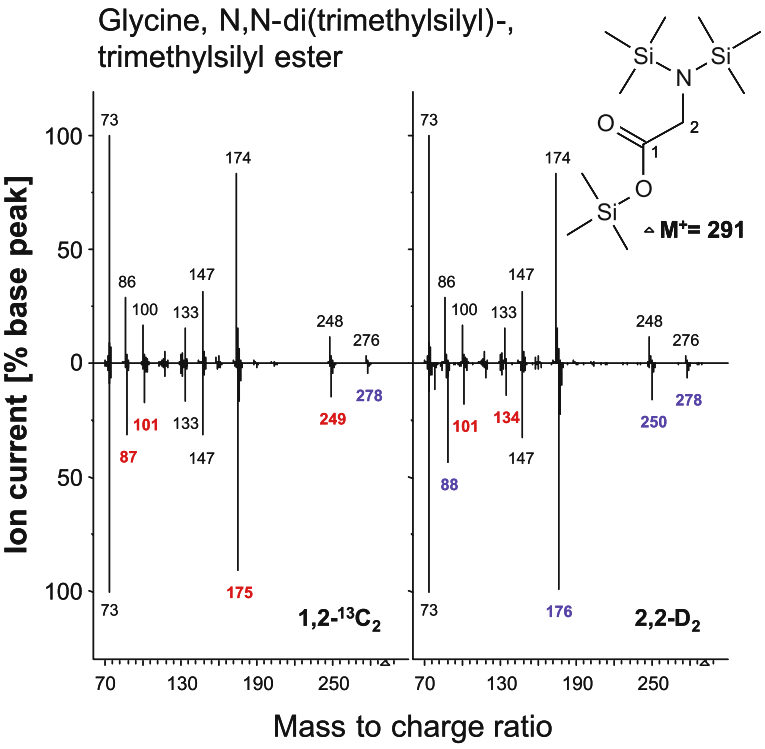
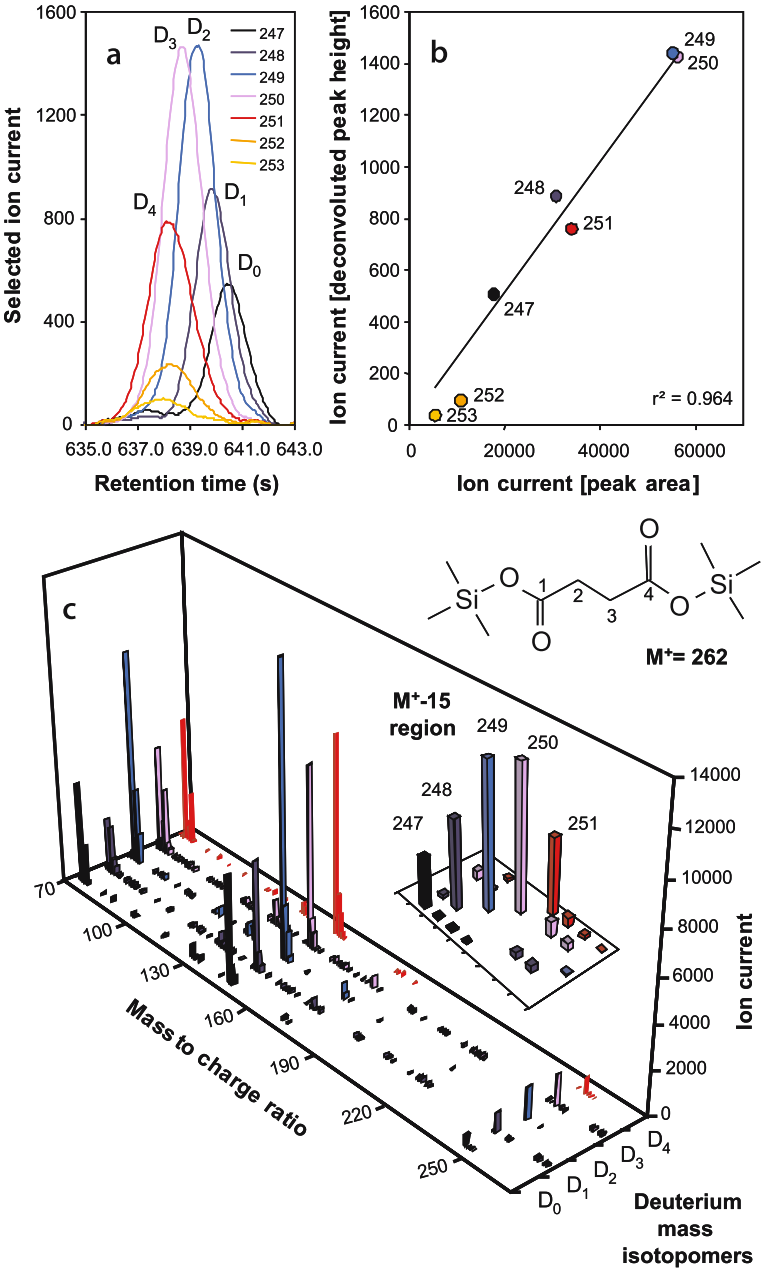


Fig. 2. Mass spectra of deuterated and ^{13}C labeled MSTs help structural elucidation and recovery analysis of metabolites. Labeled and non-labeled MSTs of Glycine *N*,*N*-di(trimethylsilyl)-, trimethylsilyl ester are shown. *Oryza sativa* L. cv. Nipponbare was labelled in vivo using deuterated water or $^{13}CO_2$. MSTs representing the fully labeled mass isotopomers demonstrate presence of two carbon atoms (*left panel*) and two non-exchangeable hydrogen atoms (*right panel*). Mass fragments which exhibited a mass shift of 1 amu (*red*) or 2 amu (*blue*) are indicated

▶ **Fig. 3a–c.** Mass spectral deconvolution of deuterated mass isotopomers. Succinic acid di(trimethylsilyl) ester was partially labelled in vivo by exposing *Oryza sativa* L. cv. Nipponbare to deuterated water. Metabolite profiles were performed on a Pegasus II GC-TOF-MS system (LECO, St. Joseph, MI, USA) with 20 scans s^{-1}. Mass spectra were deconvoluted using ChromaTOF software version 1.00, with baseline offset just above noise, smoothing and peak width set to 10 and 2 scans, respectively: **a** selected ion traces of non-deuterated (D_0, m/z = 247) and deuterated (D_{1-4}, m/z = 248 – 251) $M^+ - 15$ mass fragments. Mass fragments at 252 and 253 amu are carbon mass isotopomers of D_4; **b** peak area compared to deconvoluted peak height. Peak area integration does not allow differentiation of contributions by carbon mass isotopomers; **c** deconvoluted mass spectra of D_{0-4}. *Inset* shows partial deconvolution of D_{0-4} carbon mass isotopomers and missing carbon mass isotopomers of D_{1-3}

diverse, representative metabolites (e. g. Roessner-Tunali et al. 2003; Gullberg et al. 2004). Therefore technologies are required to improve quantitative standardisation for the comparison of increasingly diverse biological samples and experimental conditions.



For this purpose, full saturating ^{13}C in vivo labelling was developed using yeast which is one of the most important organisms in systems biology (e. g. Stephanopoulos et al. 2004). Metabolites of yeast were demonstrated to be fully labelled when provided with an exclusive carbon source, such as U-^{13}C-glucose (Mashego et al. 2004; Birkemeyer et al. 2005). Refer to Birkemeyer et al. (2005) for detailed discussion of potential applications for ^{13}C-labelled metabolomes. Similar approaches are possible in plants (Figs. 2 and 3).

In short, standardised in vivo labelled extracts of yeast or other microorganisms can substitute the rather small number of chemically synthesised mass isotopomers used in earlier studies (Fiehn et al. 2000a; Gullberg et al. 2004). Typically a standardised labelled reference sample is combined in equal amounts with non-labelled experimentally challenged samples. The advantages of this approach are (i) the presence of a mass isotopomer for all identified but also all hitherto non-identified metabolites, (ii) the concentration of each mass isotopomer is inherently adjusted to the endogenous metabolite concentration, (iii) metabolic components can easily be distinguished from laboratory contaminations, and (iv) recovery of all metabolic components can be determined with the appropriate mass isotopomer.

Thus metabolite profiling will achieve the same level of transcriptome and proteome experiments, which utilize differential fluorescent probes or differential isotope coded tagging, respectively. In conclusion, comprehensive in vivo isotope labelling will help to establish quantitative between laboratory comparability of GC-MS based metabolome experiments. More importantly, we expect metabolome experiments with full mass isotopomer standardisation to be also independent of the mass spectrometric platform, e. g. CE-MS, LC-MS, or possibly even MALDI-TOF-MS.

3 Short Excursion into Nomenclature and Definitions

Concise and unambiguous description of GC-MS metabolite profiling results requires clear definitions. The definitions suggested within this section are biased towards the specifics of GC-MS technology but may also be applied to other technology platforms. This section is intended as a contribution to the ongoing process of unifying data formats and concepts within the field of plant metabolomics (e. g. Fiehn 2002; Bino et al. 2004; Jenkins et al. 2004).

3.1 Metabolite and Analyte

Routine GC-MS profiling analysis (Fiehn et al. 2000b; Roessner et al. 2000) has an upper size exclusion limit which is roughly equivalent to a persilylated trisaccharide derivative (MW:1296), hexatriacontane (MW:506), or hentriacontanoic acid trimethylsilylester (MW:523). Even though it may appear tempting, metabolite and analyte are best not defined by molecular weight.

A *metabolite* may be described as a compound which is internalised, chemically converted or secreted by an organism, but is not synthesised by DNA replication, transcription, or translation. Post-processing events of DNA, RNA and proteins, such as DNA methylation, RNA splicing, sequence specific protease cleavage or post-translational modification are not attributed to the metabolome. The origin of a metabolite is not exclusively dependent on the biosynthetic capacity of an organism or delimited by the genomic inventory. Metabolites may readily be exchanged between organisms, for example in plant microbe interactions, and – like drugs or pesticides – can today be of anthropogenic/xenobiotic origin.

In contrast to LC- or CE-MS, GC-MS analysis requires clear distinction between metabolite and analyte, because – depending on choice of chemical derivatisation – metabolites may be chemically transformed before quantification. The term *analyte* may be used to address the chemical structure and compound which is submitted to GC-MS and finally detected and quantified. An analyte can be identical with the metabolite, if the metabolite is not chemically derivatised. Single metabolites may have more than one analyte, if the chosen derivatisation reaction generates more than one derivative, for example methoxyamination (see above). In these cases *preferred* and *alternate* analytes exist for quantification. Analytes of one metabolite may differ in abundance, i. e. a *major* and one, even multiple, *minor* analytes may exist. Standardisation by stable mass isotopomers corrects the quantification errors which may arise from unforeseen matrix effects on analyte ratios during chemical derivatisation of GC injection.

Different metabolites may be chemically transformed into the same analyte structure. In addition a single analyte may arise from inadequate chromatographic separation of isomers. For example, the biochemically distinct stereoisomeric structures of DL-amino acids are only separated by specialised chiroselective chromatography. These analytes have *composite* properties in contrast to absolutely *specific* analytes.

These concepts are not unique to GC-MS technology. Analyte sensitivity, accuracy, and potentially composite analyte properties need to be thoroughly considered in MS-MS applications, non-chiroselective capillary electrophoresis or liquid chromatography, and in cases of adduct-formation or multiply charged ions.

3.2 Mass Spectral Tag (MST) and Mass Fragment

GC-MS metabolite profiles resolve hundreds of analytes, which represent metabolites, but also internal standard substances and laboratory contaminations. Typical GC-MS profiles may contain approximately 100 identified analytes of metabolites. The chemical structure of the majority of GC-MS analytes, however, is still unknown. Each new biological object or experimental condition still gives rise to new, hitherto unidentified, chemical components.

Because in non-biased analysis of GC-MS profiles identified and unidentified components are equally important, we created the term *mass spectral tag* (MST), i. e. a mass spectrum which is characterised by a specific chromatographic retention and by repeated occurrence in a single or multiple types of biological samples (Colebatch et al. 2004; Desbrosses et al. 2005). MSTs represent analytes. MSTs can be identified, in other words, unequivocally linked to a chemical structure. The use of MSTs allows uncoupling of metabolite profiling experiments from the time consuming process of chemical identification. MSTs can be used to track analytes in different experiments or laboratories (Schauer et al. 2005). Thus MST identification can be performed even years after the first discovery.

MSTs of GC-EI-MS profiles are composed of multiple characteristic *mass fragments* in constant relative abundances. In most cases residual, non-fragmented molecular ions are rare or even absent. In consequence GC-MS allows selection of multiple mass fragments which all represent the same MST and exhibit the same quantitative changes. Typically one quantifying mass fragment (QM) and a set of specific, supporting qualifying mass fragments are selected in GC-MS analysis (Halket et al. 2005). The criteria for the proper choice of QMs are equal to the choice of a preferred analyte. QMs need to be selective, i. e. not composite, in the context of the complexity of co-eluting MSTs. Therefore, the best QM is the most abundant among the available selective mass fragments.

3.3 Response and Relative Quantification of Metabolite Pools

GC-MS metabolite profiling studies monitor relative changes in metabolite pool sizes and but also allow insight into flux, i. e. the dynamic turnover of metabolite pools or metabolite substructures (e. g. Fischer and Sauer 2003; Sauer 2004; Roessner-Tunali et al. 2004). Flux experiments are easily distinguished from above mentioned saturating in vivo labelling experiments. Flux experiments monitor the initial kinetics of labelling and thus stable isotopes are only partially incorporated into metabolite pools. In contrast, saturating in vivo labelling reaches the endpoint of a completely stable isotope labelled metabolome.

MSTs are quantified by ion currents of QMs which are recorded after analyte ionization, fragmentation and mass separation. Ion currents in GC-MS are monitored either by peak area or peak height. Both measurements need to be baseline corrected for electronic and chemical noise. The resulting corrected values are defined to be what we call *responses*, i. e. X_{QM} of fragment QM (Colebatch et al. 2004; Desbrosses et al. 2005). The fragment response is routinely normalised to the amount of the sample, for example fresh or dry weight. In addition each response is corrected for recovery effects, which may occur at any step of the analytical process between metabolic inactivation of the sample and final recording of ion currents. Different levels of recovery correction exist: (i) correction by extract and sample volume, (ii) correction

by addition of a constant amount of a representative internal standard compound (IS), and (iii) normalization by chemically identical, but stable-isotope labelled mass isotopomers of each metabolite. The *normalised response* (N_{QM}) is, consequently, $N_{QM} = X_{QM} \times X_{IS}^{-1} \times \text{sample weight}^{-1}$, where X_{IS} ideally represents a mass isotopomer response of QM. In a further step, the normalised response of a fragment, N_{QM}, is divided by the average relative response of QM as determined in a set of reference samples, $\text{avg}N_{QM(ref)}$. The resulting quotient, $R_i = N_{QM} \times \text{avg}N_{QM(ref)}^{-1}$, is called *response ratio* R_i. R_i describes the x-fold changes in metabolite pools sizes relative to the reference samples. Typical reference samples are taken at the start of a time series experiment or are mock-treated biological controls.

In GC-MS profiling analyses the standard deviation of normalised responses is dependent on the chemical nature of metabolite and analyte. Average relative standard deviations (RSD) of 10% (Weckwerth et al. 2004b) or 13.8% (5.5–33.4%; Gullberg et al. 2004) were reported for replicate GC-MS analyses. These analyses included extraction as well as derivatisation and were performed using representative analytes. Use of isotope labelled standardisation was reported to reduce RSD further to approximately 6.9–9.7% residual experimental variance (Gullberg et al. 2004).

4 Present Challenges of GC-MS Profiling

4.1 Standardisation of GC-MS Systems

GC-MS profiles, with the exception of GC×GC-TOF-MS data, are in essence three-dimensional and comprise a chromatographic time-resolved axis, a second coordinate axis which represents the mass to charge ratio (m/z, z = 1 in GC-MS with only rare exceptions), and an intensity axis which monitors the ion current (IC) and thus the abundance of molecules or mass fragments. A substantial breakthrough for GC-MS analyses was the early establishment of generally accepted calibration substances and procedures, so-called tuning routines, which allowed comparison of mass spectra from GC-MS systems of virtually all manufacturers and from different hyphenated mass detection technologies. In addition the widely used electron-impact ionisation technique (EI) ensured stable fragmentation ratios, which are in first approximation independent of analyte concentration. However, comparability was only achieved by restriction to 1 amu precision.

The chromatography axis is less standardised, not least because of multiple types of available capillary GC-columns which have different chromatographic properties and thus serve different separation problems. In addition slight changes in temperature program, pressure and flow settings of both carrier gas and injection technique, as well as slight production differences of column manufacturers cause minor but perceptible changes in retention

time. Retention time indices (RI), based on homologous series of internal reference substances, such as *n*-alkanes, have been introduced early to aid GC analyses (Kovàts 1958). Use of an *n*-alkane RI system in GC-MS metabolite profiling substantially improves the reproducibility of the chromatography axis. The currently achievable accuracy of RI prediction was recently investigated in three different profiling laboratories which use the same type of capillary column but different GC-MS systems (Schauer et al. 2005). In this investigation the possibility of predicting RIs of more than 100 identified analytes was tested. Mathematical regression resulted in an average accuracy of ±5.4 RI units.

The IC intensity axis in GC-MS is standardised for high vs low mass discrimination. The GC-MS tuning includes processes which ensure constant ratios of high vs low mass intensities. However, mass spectra which are recorded by either QUAD-MS or fast scanning GC-TOF-MS detection may differ in this respect. Fast scanning GC-TOF-MS systems (e. g. Pegasus II MS system, LECO, St. Joseph, MI, USA) have increased sensitivity of small mass fragments and reduced sensitivity in the high mass range.

4.2 Deconvolution and Alignment of Mass Spectral Tags

The principal challenge in GC-MS profiling analysis is the automated unravelling of the multiple partially co-eluting MSTs which comprises a GC-MS chromatogram. One of the fundamental advances in GC-MS technology has been the development of algorithms and software for the so-called deconvolution of mass spectra from GC-MS chromatograms (Halket et al. 1999; Stein 1999; Shao et al. 2004; AnalyzerPro, http://www.spectralworks.com), specific software for the deconvolution of fast scanning GC-TOF-MS data files, e. g. ChromaTOF software used by Vreuls et al. (1999), Veriotti and Sacks (2001) or Jonsson et al. (2005), and ongoing developments for automated processing of GC×GC-TOF-MS chromatograms (Ryan et al. 2004; Sinha et al. 2004a,b). The inherent steps of deconvolution are (i) mass resolved baseline subtraction of electronic and chemical noise, (ii) assignment of retention time and/or retention time indices (RI) to chromatographic peak apices and respective MSTs, and (iii) accurate separation of MSTs from closely co-eluting analytes, the most challenging and advanced but still error-prone part (Fig. 3).

Even though automated mass spectral deconvolution has fundamentally facilitated GC-MS analyses of complex mixtures, accuracy and limitations of respective software have so far not been systematically compared and assessed. Typical errors of mass spectral deconvolution are (i) accidental generation of MSTs due to noise fluctuations, (ii) deconvolution of multiple MSTs from a single component, (iii) incomplete MSTs which lack one or multiple mass fragments (Fig. 3c), and (iv) chimeric MSTs, i. e. composite mass spectra of co-eluting compounds. The co-elution problem of complex mixtures has been fundamentally improved by introduction of fast scanning GC-TOF-MS and is

today technically best solved by GC×GC-TOF-MS, using a set of two capillary GC columns with alternate phase-polarity (Sinha et al. 2004c).

Reliable alignment of identical MSTs in sets of consecutive GC-MS chromatograms is required for rapid, repeatable and automated comparative high-throughput analysis of large samples sets. So far, software solutions and novel algorithm developments for the alignment of complex mixtures depend on close to constant chromatographic retention within series of consecutive GC-MS chromatograms (Duran et al. 2003; Jonsson et al. 2004; metAlign, http://www.metalign.nl). Indeed, consistent run-to-run retention times are considered to be crucial to the application of chemometrics on complex mixtures, especially in the field of two-dimensional separations (Sinha et al. 2004c).

In conclusion, automated mass spectral deconvolution of GC-MS profiles appears to be in principal solved by both GC×GC technology and deconvolution algorithm, but the optimum solution still has to be found (Halket et al. 2005). In contrast prediction of chromatographic shifts in complex mixtures with highly dynamic range of concentrations has not been satisfyingly solved. As there is currently no solution – other than recalibration with pure standard substances – addressing the problem of RI shifts will be crucial for future GC-MS based metabolite profiling and identification of MSTs.

4.3 Identification of Mass Spectral Tags

Identification of MSTs requires chromatographic separation as well as mass spectrometric information (Wagner et al. 2003), mainly because plants like microorganisms contain a multitude of isomeric metabolites (e. g. Barsch et al. 2004; Stephanopoulos et al. 2004; Strelkov et al. 2004). These isomers give rise to MSTs, which can be chromatographically resolved but have almost identical mass spectra. Today, GC-MS appears to have found a generally accepted standard for mass spectral comparison. The NIST mass spectral search and comparison software (Ausloos et al. 1999; Stein 1999) has been integrated into the customised operating software of most GC-MS manufacturers. The GC-MS technology is in this respect more advanced than LC-MS (Halket et al. 2005). However, mass spectral search and comparison software, which harbours information on chromatographic retention in what we suggested to call MSRI libraries (Wagner et al. 2003; Kopka et al. 2005) and the automated utilization of this information for probability based matching, would be highly desirable. The new version NIST05 (National Institute of Standards and Technology, Gaithersburg, MD, USA) of a mass spectral search and comparison software now makes RI information available but currently does not utilize RI for automated matching. The result of a hitherto manual inventory of *Oryza sativa* L. cv. Nipponbare leave profiles is shown in Fig. 4.

Two different approaches exist for the identification of unknown MSTs from GC-MS profiles: (i) the “bottom up” approach in which metabolites of interest to a particular researcher are analysed by the purchase of authentic standard

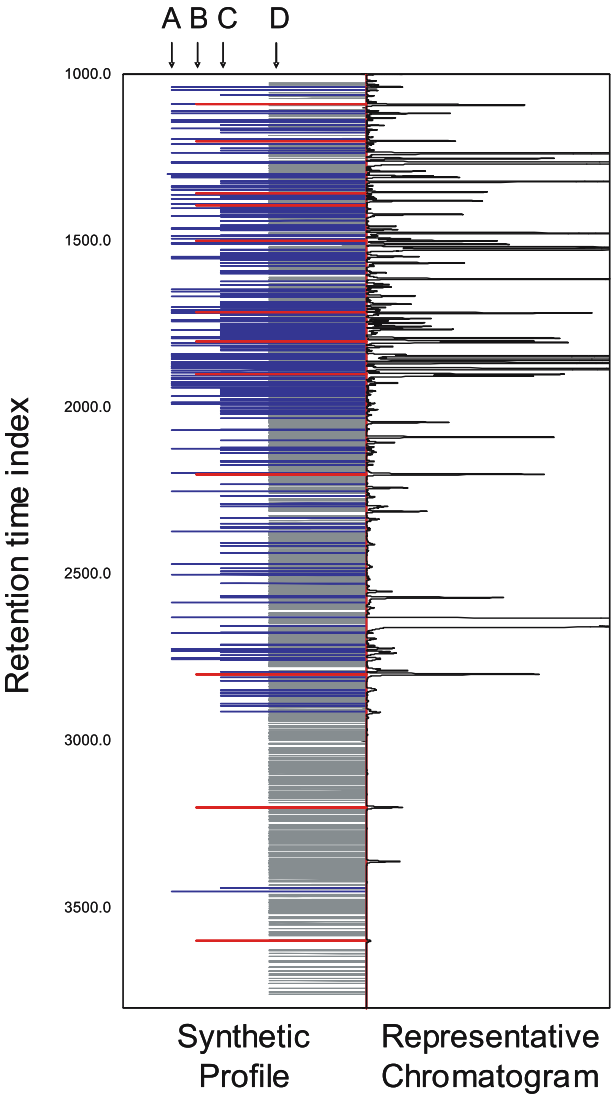


Fig. 4. Synthetic and representative GC-MS profiles of *Oryza sativa* L. cv. Nipponbare leaves: A – 132 identified MSTs representing 109 known metabolites; B – 12 added internal standard substances; C – 148 unidentified MSTs which match previous MSRI library entries; D – all previously observed MSTs present in the MSRI library at GMD (http://csbdb.mpimp-golm.mpg.de/gmd.html)

substances, which are subsequently mapped by standard addition experiments onto established standardised GC-MS systems, and (ii) the "top down" approach whereby structural elucidation is performed on a hitherto unknown, but important target MST. The work of "top down" structural identification

is highly time-demanding and involves preparative enrichment, purification, spectroscopic, mass spectral and NMR analyses of the preparation and finally chemical synthesis of the predicted structure. As a consequence the "bottom up" approach prevails in most laboratories and "top down" identification is currently restricted to potentially novel signalling compounds or marker substances of specific biological samples and experimental conditions.

In order to avoid unnecessary "top down" investigations reliable identification by prior standard addition experiments is essential. MSRI library collections of mass spectra (Kopka et al. 2005), which comprise frequently observed identified and non-identified MSTs, appear to represent the most effective means to pool the identification efforts currently performed in many laboratories around the world (Schauer et al. 2005). Identified and yet unidentified, MSTs can efficiently be shared by public resources such as GMD@CSBDB (http://csbdb.mpimp-golm.mpg.de/gmd.html). In addition mass spectral identifications and chromatographic sequence of analyte elution can be transferred between laboratories. "Bottom up" identifications performed in parallel may be used for inter-laboratory confirmation of identifications and reduce the risk of unnecessary structural elucidation projects.

Currently the MSRI libraries available from GMD@CSBDB include in total more than 2000 evaluated mass spectral data sets obtained using GC-QUAD- and GC-TOF-MS technology platforms with 1089 non-redundant and 360 identified MSTs. Future efforts at GMD aim to refine mass spectral quality and annotation, and will add stable isotope labelled variants of MSTs (e. g. Fig. 2) for improved mass spectral interpretation of unidentified MSTs. The number of identified analytes and metabolites will continuously be extended and annotations updated.

Acknowledgements. I would like to thank A.R. Fernie, A. Erban, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany for critically reading and discussing my manuscript. My thanks extend to Prof. Dr. Le Tran Binh, Institute of Biotechnology (IBT), Hanoi, Vietnam, for sharing his expertise in rice cultivation. This work was supported by the Max-Planck society, and the Bundesministerium für Bildung und Forschung (BMBF), grant PTJ-BIO/0312854.

References

Ausloos P, Clifton CL, Lias SG, Mikaya AI, Stein SE, Tchekhovskoi DV, Sparkman OD, Zaikin V, Zhu D (1999) The critical evaluation of a comprehensive mass spectral library. J Am Soc Mass Spectrom 10:287–299

Barsch A, Patschkowski T, Niehaus K (2004) Comprehensive metabolite profiling of *Sinorhizobium meliloti* using gas chromatography-mass spectrometry. Funct Integrat Genomics 4:219–230

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci 9:418–425

Birkemeyer C, Kolasa A, Kopka J (2003) Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. J Chromatogr A 993:89–102

Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J (2005) Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling. Trends Biotechnol 23:28–33

Blau K, Halket JM (1993) Handbook of derivatives for chromatography, 2nd edn. Wiley, New York

Broeckling CD, Huhman DV, Farag MA, Smith JT,May GD, Mendes P, Dixon RA, Sumner LW (2005) Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. J Exp Bot 56:323–336

Colebatch G, Desbrosses G, Ott T, Krusell L, Kloska S, Kopka J, Udvardi MK (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. Plant J 39:487–512

Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. Proc Natl Acad Sci USA 101:15243–15248

Desbrosses G, Kopka J, Udvardi MK (2005) Legume metabolomics: development of GC-MS resources for functional genomics of plant-microbe interactions. Plant Physiol 137:1302–1318

Duran AL, Yang J, Wang L, Sumner LW (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). Bioinformatics 19:2283–2293

Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5:763–769

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48:155–171

Fiehn O (2003) Metabolic networks of *Cucurbita maxima* phloem. Phytochem 62:875–86

Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000a) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal Chem 72:3573–3580

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000b) Metabolite profiling for plant functional genomics. Nat Biotechnol 18:1157–1161

Fischer E, Sauer U (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. Eur J Biochem 270:880–891

Gullberg J, Jonsson P, Nordström A, Sjöström M, Moritz T (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. Anal Biochem 331:283–295

Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids – potential for pattern recognition and automated identification of metabolic disorders. Rapid Commun Mass Spectrom 13:279–284

Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. J Exp Bot 56:219–243

Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall RD, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nature Biotechnol 22:1601–1606

Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, Moritz T (2004) A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem 76:1738–1745

Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, Marklund S, Sjostrom M, Antti H, Moritz T (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. Anal Chem 77: 5635–5642

Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of *Arabidopsis*. Plant Physiol 136:4159–4168

Knapp DR (1979) Handbook of analytical derivatization reactions. Wiley, New York
Kopka J, Fernie AF, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in Plant Biology: Platforms and Destinations. Genome Biol 5(6):109–117
Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSBDB: The Golm Metabolome Database. Bioinformatics 21:1635–1638
Kovàts ES (1958) Gas-chromatographische Charakterisierung organischer Verbindungen: Teil 1. Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helv Chim Acta 41:1915–1932
Little JL (1999) Artifacts in trimethylsilyl derivatization reactions and ways to avoid them. J Chromatorgr A 844:1–22
Mashego MR, Wu L, van Dam JC, Ras C, Vinke JL, van Winden WA, van Gulik WM, Heijnen JJ (2004) MIRACLE: mass isotopomer ratio analysis of U-^{13}C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. Biotech Bioeng 85:620–628
Mueller A, Duechting P, Weiler EW (2002) A multiplex GC-MS/MS technique for the sensitive and quantitative single-run analysis of acidic phytohormones and related compounds, and its application to *Arabidopsis thaliana*. Planta 216:44–56
Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23:131–142
Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001a) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13:11–29
Roessner U, Willmitzer L, Fernie AR (2001b) High-resolution metabolic phenotyping of genetically and environmentally diverse plant systems – identification of phenocopies. Plant Physiol 127:749–764
Roessner U, Willmitzer L, Fernie AR (2002) Metabolic profiling and biochemical phenotyping of plant systems. Plant Cell Rep 21:189–196
Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol 133:84–99
Roessner-Tunali U, Lui J, Leisse A, Balbo I, Perez-Melis A, Willmitzer L, Fernie AR (2004) Flux analysis of organic and amino acid metabolism in potato tubers by gas chromatography-mass spectrometry following incubation in ^{13}C labelled isotopes. Plant J 39:668–679
Ryan D, Shellie R, Tranchida P, Casilli A, Mondello L, Marriott P (2004) Analysis of roasted coffee bean volatiles by using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. J Chromatogr A 1054:57–65
Sauer U (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol 15:58–63
Sauter H, Lauer M, Fritsch H (1988) Metabolite profiling of plants – a new diagnostic technique. Abstr Pap Am Chem Soc 195:129
Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579:1332–1337
Schmelz EA, Engelberth J, Alborn HT, O'Donnell P, Sammons M, Toshima H, Tumlinson JH (2003) Simultaneous analysis of phytohormones, phytotoxins, and volatile organic compounds in plants. Proc Natl Acad Sci USA 100:10552–10557
Schmelz EA, Engelberth J, Tumlinson JH, Block A, Alborn HT (2004) The use of vapor phase extraction in metabolic profiling of phytohormones and other metabolites. Plant J 39:790–808
Shao XG, Wang GQ, Wang SF, Su QD (2004) Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background. Anal Chem 76:5143–5148

Sinha AE, Fraga CG, Prazen BJ, Synovec RE (2004a) Trilinear chemometric analysis of twodimensional comprehensive gas chromatography-time-of-flight mass spectrometry data. J Chromatogr A 1027:269–277

Sinha AE, Hope JL, Prazen BJ, Nilsson EJ, Jack RM, Synovec RE (2004b) Algorithm for locating analytes of interest based on mass spectral similarity in GC×GC-TOF-MS data: analysis of metabolites in human infant urine. J Chromatogr. A 1058:209–215

Sinha AE, Prazen BJ, Synovec RE (2004c) Trends in chemometric analysis of comprehensive two-dimensional separations. Anal Bioanal Chem 378:1948–1951

Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J Am Soc Mass Spectrom 10:770–781

Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J (2004) CSB.DB: a comprehensive systems-biology database. Bioinformatics 20:3647–3651

Stephanopoulos G, Alper H, Moxley J (2004) Exploiting biological complexity for strain improvement through systems biology. Nat Biotechnol 22:1261–1267

Strelkov S, von Elstermann M, Schomburg D (2004) Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. Biol Chem 385:853–861

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Toyo'oka T (1999) Modern derivatization methods for separation science. Wiley, New York

Trethewey RN (2004) Metabolite profiling as an aid to metabolic engineering in plants. Curr Opin Plant Biol 7:196–201

Trethewey RN, Krotzky AJ, Willmitzer L (1999) Metabolic profiling: a Rosetta stone for genomics? Curr Opin Plant Biol 2:83–85

Urbanczyk-Wochniak E, Fernie AR (2005) Metabolic profiling reveals altered nitrogen nutrient regimes have diverse effects on the metabolism of hydroponically-grown tomato (*Solanum lycopersicum*) plants. J Exp Bot 56:309–321

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO Reports 4:989–993

Van Deursen MM, Beens J, Janssen HG, Leclercq PA, Cramers CA (2000) Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. J Chromatogr A 878:205–213

Veriotti T, Sacks R (2001) High-speed GC and GC/time-of-flight MS of lemon and lime oil samples. Anal Chem 73:4395–4402

Vreuls RJJ, Dallüge J, Brinkman UAT (1999) Gas chromatography – time-of-flight mass spectrometry for sensitive determination of organic microcontaminants. J Microcolumn Sep 11:663–675

Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochem 62:887–900

Weckwerth W, Loureiro ME, Wenzel K, Fiehn O (2004a) Differential metabolic networks unravel the effects of silent plant phenotypes. Proc Natl Acad Sci USA 18:7809–7814

Weckwerth W, Wenzel K, Fiehn O (2004b) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. Proteomics 4:78–83

I.2 Current Status and Forward Looking Thoughts on LC/MS Metabolomics

L.W. Sumner[1]

1 Introduction

The metabolome can be viewed as the consequential end products of gene expression and the goal of metabolomics includes the comprehensive evaluation of the metabolome (Trethewey et al. 1999; Fiehn et al. 2000; Trethewey 2001; Oliver et al. 2002; Sumner et al. 2003). Quantitative and qualitative measurements of large numbers of cellular metabolites provide a high-resolution biochemical phenotype of an organism which can be used to monitor and assess gene function (Fiehn et al. 2000) or a system's response (Weckwerth 2003). Although mRNA/transcripts represent a mechanism for information transmission from the genome to the cellular machinery for protein synthesis, mRNA levels do not always correlate well with protein levels (Gygi et al. 1999). Furthermore, once translated a protein may or may not be enzymatically active as post translational modifications, protein sorting, protein–protein interactions, and controlled proteolysis all contribute to the regulation of active enzyme levels. Due to these factors, changes in the transcriptome or the proteome may not always lead to alterations in the metabolic phenotype. In addition, the majority of transcript and protein annotations are currently inferred based on sequence or structural similarity. It is estimated that less than 10% of annotated genes have experimental evidence supporting assigned function and thus, the accuracy of these annotations are of some uncertainty (Somerville and Somerville 1999; Somerville and Dangl 2000). In the absence of functionally annotated database information, transcript or protein profiling often yields limited information. For example, transcriptomics or proteomics often reveal the differential accumulation of a hypothetical or unannotated protein; however, without annotation it is very difficult to infer biological context. Microarray or proteomics experiments may also yield putative or generic protein identifications such as a putative peroxidase or peroxidase-like protein. These generic annotations have limited information as many of these enzymes are promiscuous and/or involved in a large number of different reactions. However, metabolomics has the ability to reveal that the accumulated peroxidase/enzyme is more specifically related to lignification or to another specific biochemistry. Thus, profiling the metabolome may actually provide the most direct and "functional" information of the "omics" technologies.

[1] The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA, e-mail: lwsumner@noble.org

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

The plant metabolome is quite complex with current estimates on the order of 15,000 metabolites within a given species and over 200,000 different metabolites within the plant kingdom (Dixon 2001; Hartman et al. 2005). Due to the chemical complexity of the plant metabolome, it is generally accepted that a single analytical technique will not provide comprehensive visualization of the metabolome, and therefore, multiple technologies are generally employed. The selection of the most suitable technology is generally a compromise between speed, chemical selectivity and instrumental sensitivity. Tools such as nuclear magnetic resonance spectroscopy (NMR) are rapid, highly selective, and non-destructive, but have relatively lower sensitivity. Other methods such as capillary electrophoresis (CE) coupled to laser induced fluorescence (LIF) detection are highly sensitive, but have limited chemical selectivity. Chromatographically coupled mass spectrometry methods such as gas chromatography (GC)/mass spectrometry (MS) and liquid chromatography (LC)/MS offer the best combination of sensitivity and selectivity, and therefore are central to most metabolomics approaches. Mass selective detection provides highly specific chemical information including molecular mass and/or characteristic fragment ion(s) information that are directly related to chemical structure. This information can be utilized for compound identification through spectral matching with data compiled in libraries for authentic compounds or used for de novo structural elucidation. Further, chemically selective MS information can be obtained from extremely small metabolite quantities with limits of detection in the pmole and fmole level for many primary and secondary plant metabolites.

GC/MS has proven capability for profiling large numbers of metabolites with reports covering several hundred to slightly more than a thousand various components (Fiehn et al. 2000; Roessner et al. 2000, 2001; Birkemeyer et al. 2003; Wagner et al. 2003; Broeckling et al. 2005; Schauer et al. 2005; Welthagen et al. 2005). The term component is used because a large number of metabolites often yield more than one derivatized component which are observed in the GC/MS analysis. The achievable range and number of metabolites profiled by GC/MS can be attributed to the high separation efficiencies of long (30−60 m) capillary GC columns (i. e. $N \geq 250{,}000$ for 60 m). These high efficiencies enable the separation of very complex mixtures, and with mass selective detection, qualitative identification of a significant proportion of these compounds is achievable. This makes GC/MS a very efficient and cost effective metabolomics tool. A major prerequisite for GC/MS is sample volatility which is necessary to enable separation in the gas phase. Analytes may be innately volatile or chemically derivatized to yield volatile compounds. Unfortunately, there exist a large number of metabolites which are not amenable to GC/MS even following derivatization. These include compounds such as phenylpropanoid and other natural product glycosides whose labile glycosidic bonds degrade during heating and vaporization. Thus, alternative techniques are necessary and especially so for the study of secondary metabolism.

Liquid introduction techniques for mass spectrometry such as electrospray ionization, atmospheric chemical ionization, and photo ionization remove the necessity for chemical derivatization. Thus, aqueous samples can be analyzed with minimal sample processing or even directly from the tissue source (Takats et al. 2004). Further, these techniques allow for the analyses of more labile and larger metabolites, and for the coupling of liquid separation technologies to mass spectrometry. Therefore high performance liquid chromatography (HPLC) and CE are readily coupled to mass spectrometry to yield powerful tools for targeted metabolic profiling and non-targeted metabolomics.

The utility of LC/MS emanates from the coupling of a 'universal' separation technology to a selective and sensitive mass analyzer detector. HPLC is commonly considered a universal separation technique because of its applicability to a broad range of chemical classes with a diversity of physical and chemical properties. For example, HPLC has been utilized for the analysis of ionic compounds, inorganics, volatile organics, polar organics, non-polar organics, lipids, amino acids, carbohydrates, nucleotides, carotenoids, phenylpropanoids, hormones, peptides, proteins, and the list goes on. The major point is that HPLC can be used for many of those compounds commonly analyzed by GC and many more. LC/MS also removes the need for derivatization and thus, complex samples can be analyzed directly or with minimal sample processing. As a result of these favorable properties, it is not surprising that LC/MS and LC coupled to tandem mass spectrometry (LC/MS/MS) have become popular tools for metabolism investigations.

HPLC is performed on various scales utilizing different column sizes. General values are provided in Table 1 for preparative, analytical, micro, capillary and nano-scale modes of HPLC. Generally, preparative scale HPLC is used for compound(s) purification and analytical scale is traditionally used for the quantitative analyses of plant extracts. However, smaller scale technologies (micro, capillary, nano) are now commercially available for quantitative analyses. These smaller scale separations offer significant sensitivity enhancements, and thus reduce the amount of material necessary for analysis. Further, capillary and nano HPLC often offer increased chromatographic resolution. Unfortunately as the separation scale gets smaller it becomes more difficult to reproducibly generate mobile phase gradients and the retention time variance increases. However, this problem is continually decreasing as novel instrumentation and approaches become available.

Table 1. General liquid chromatographic scales

Scale	Column internal diameter	Flow rate
Preparative	2.1–>200 mm	10 mL/min,
Analytical (*conventional*)	2.1–4.6 mm	1.0 mL/min
Micro	1.0 mm	200 µL/min
Capillary	300 µm–1 mm id	4 µL/min
Nano	25–300 µm id	200 nL/min

2 Chromatography Theory

Currently, the chromatographic performance of HPLC, relative to GC and CE, is lower, and there is a significant need for improvement. However, to discuss this issue and possible improvements in detail, several terms must be defined. A number of quantifiers are used to assess chromatographic performance. These include resolution (R_s), selectivity (α), efficiency (N), and peak capacity (n) which are defined below:

1. *Resolution* (R_s) is a quantifier of the degree of separation between mixture components, i. e. two peaks t_a and t_b with peak widths at the base w_a and w_b. A resolution of 1 indicates that two adjacent peaks are baseline resolved. Resolution can also be expressed as a function of the theoretical plate number (N) and selectivity (α) as defined below in Eq. (1):

$$R_s = \frac{2(t_b - t_a)}{w_a + w_b} = \frac{2\Delta t_R}{w_a + w_b} \qquad R_s = \frac{\sqrt{N}}{4}\left(\frac{\alpha - 1}{\alpha}\right)\left(\frac{k_2}{1 + k_2}\right) \tag{1}$$

2. *Selectivity* (α), which is also referred to as the separation factor, is a ratio of the retention or capacity factor (k') of two peaks. The capacity factor is a relative retention parameter that has been normalized using the void elution time (t_v) or volume (V_v) and is therefore independent of column geometry – see Eq. (2). The void value is the volume or time of an unretained component. The selectivity parameter provides a quantifier of the relative separation of two components. Selectivity can be altered based on the chemical composition of the stationary phase, stationary phase manufacturer, mobile phase, and pH:

$$\alpha = \frac{k'_2}{k'_1} \qquad k'_1 = \frac{t_1 - t_v}{t_v} = \frac{V_1 - V_2}{V_v} \tag{2}$$

3. *Column efficiency* is usually quantified based upon a column's theoretical plate number (N) which is unitless and a measure of band broadening per unit time – see Eq. (3). This can be practically quantified using retention time (t_R) and peak width. Peak width can be defined at the base (W_b) or at half height ($w_{1/2}$) as they are directly related if one assumes a Gaussian peak shape, i. e. $W_b = 1.698\ w_{1/2} = 4\sigma$ where σ equals the standard deviation of the peak. Alternatively, plate number can be calculated using the column resolution (R) and selectivity (α).
4. *Separation efficiency* is also quantified using a normalized theoretical plate number based on column length, i. e. (N/L) with units of plates/m. The theoretical plate number can be dramatically increased by decreasing the peak width. Plate number and efficiency are also related to particle size (d_p) and column length (L) as described below:

$$N = \left(\frac{t'_R}{\sigma}\right)^2 = 16\left(\frac{t'_R}{W_b}\right)^2 = 5.54\left(\frac{t'_R}{w_{1/2}}\right) = \frac{16R^2}{(1 - \alpha)^2} = \frac{L}{d_p} \tag{3}$$

5. *Peak capacity* (n) is a measure of the maximum number of theoretical peaks resolvable by the chromatographic system based on optimum performance and equal variation in the partitioning of all components in the mixture – see Eq. (4). The peak capacity is a good parameter for estimating the maximum number of compounds resolvable by a given chromatographic system. Ideally this value should approach or exceed the number of compounds that need to be separated, i. e. the number of metabolites:

$$n = \frac{\sqrt{N}}{4R} \ln\left(\frac{t_2}{t_1}\right) + 1 \tag{4}$$

3 Limitations of Current Metabolic Profiling Approaches and Proposed Solutions to Advance Metabolomics

Currently, the major limitation of metabolomics is its inability to comprehensively profile all of the metabolome. This inability is directly related to the chemical complexity of the metabolome, the biological variance inherent in most living organisms, and the dynamic range limitations of most instrumental approaches (Sumner et al. 2003). Many biological responses to altered gene expression or to environmental stimuli result in both quantitative and qualitative changes in metabolite pools. Understanding these responses is most dependent upon the qualitative identification of the altered metabolite. Quantitative measurements are also important, as both temporal and spatial changes in metabolite concentrations are expected; however this information is of little use if it cannot be assigned to a specific metabolite or biological process. Thus, comprehensive qualitative and quantitative analysis of all metabolites within a cell, tissue or organ is the ultimate goal of metabolomics; however, this is still a very ambitious goal and far from a reality for any system. Bino and colleagues (Bino et al. 2004) proposed two major objectives to increase the comprehensive nature of metabolomics. They were:

1. Increase the current capacity for metabolite separation and differentiation (i. e. the number of resolvable components within the complex metabolome mixture) using multi-dimensional separations.
2. Increase the number of identifiable metabolites through the generation of spectral libraries, high resolution accurate mass measurements, and tandem mass spectrometry.

Unfortunately, the separation of complex metabolome mixtures is still quite challenging. Currently, analytical scale HPLC (4.6 × 250 mm) is most commonly used for natural product analyses; however, the upper peak capacities (i. e. theoretical number of maximum peaks resolvable based on optimum performance) of these systems is approximately 300 (Tanaka et al. 2004). Based on this estimate, a maximum of 300 components could be resolved in a best

case scenario; however in practice this value is seldom achieved and more realistic peak capacities are between 100 and 200. Thus, current HPLC technologies are limiting the comprehensive scope of metabolomics. Separation efficiencies can be improved by altering selectivity, increasing column lengths, reducing particle sizes, increasing temperature, and/or alternative column materials. Alternatively, the utilization of multidimensional chromatography offers increased HPLC peak capacities of greater than 1000 to provide more comprehensive coverage of plant natural products (Tanaka et al. 2004). Each of methods to increase HPLC efficiency is discussed below.

Typically, improving selectivity is the best approach to improving chromatographic resolution. Selectivity is based upon the chemical or physical interaction properties that are fundamental to the separation process. More precisely, the separation selectivity of specific components can be optimized by the appropriate choice of column materials, mobile phases, and/or manufacturer. Various generic and proprietary materials are available for various chromatographic modes for HPLC. Example modes include ion-exchange, normal-phase, reverse-phase, hydrophilic interaction, and size exclusion chromatography. All HPLC columns are not equal, and different particles, particle sizes, surface modification chemistries, surface coverage, and packing processes vary significantly from manufacturer to manufacturer. These parameters dramatically influence chromatographic performance.

Often selectivity is optimized for a targeted set of analytes as a means of increasing resolution. However, in more complex mixtures associated with global metabolomics-based approaches, improved selectivity for one class of compounds often results in decreased selectivity for others. Thus, techniques (e. g. reverse-phase chromatography) with a broad range of selectivity are most likely to be the best choices for metabolomics.

One of the simplest means of increasing resolution is to increase the number of theoretical plates. Since the plate number is directly proportional to the column length (Eq. (3)), one needs only to increase the column length to increase resolution. However, Eq. (1) tells us that R is proportional to the square root of N. Thus, to achieve a 2× increase in resolution, we would have to square the column length. For example a 250 mm long column would need to be extended to 625 cm (i. e. 25 × 25 cm) for a twofold increase in resolution. Unfortunately, this is not a practical solution as the operating pressure is directly proportional to the column length. Equation (5) defines the relationship between pressure (ΔP), column length (L), analyte diffusion coefficient (D_m), particle size (d_p), mobile-phase viscosity (η), and column permeability (K^o):

$$\Delta P = \left(\frac{L v D_m}{d_p}\right)\frac{\eta}{K^o} \tag{5}$$

If a typical column of 25 cm has an operational pressure of 3000 pounds per square inch (p.s.i.), then a twofold resolution increase obtained by squaring the

column length $(25\ \text{cm})^2$ would require an operational pressure of 75,000 p.s.i. (i. e. 3,000 p.s.i. × 25). Although this illustrates the advantage of very high pressure liquid chromatography which has been achieved by select groups using custom apparatuses (MacNair et al. 1997, 1999; Tolley et al. 2001; Patel et al. 2004; Shen et al. 2005), commercial pumps do not operate at these pressures (most commercial HPLC pumps have a 5,000-p.s.i. limit). Therefore, significant resolution enhancements achieved through longer columns is limited for most researchers. With that said, several companies (i. e. Waters and JASCO) have recently introduced 15,000-p.s.i. HPLC pumps.

Equation (5) reveals that the pressure differential is proportional to the mobile phase viscosity (η). Thus, lowering of the mobile phase viscosity (η) by increasing the temperature can lower the operational pressure and allow the use of longer columns for resolution enhancement (Djordjevic et al. 1998, 1999, 2000). Selectivity is also affected by temperature and additional efficiency can be achieved by heating alone. However, one must ensure analyte thermal stability if elevated temperature separations are to be employed.

Equation (5) also shows that the pressure is a function of the column permeability (K^o). New monolithic columns offer greater permeability and lower pressures, thus allowing for the use of longer columns. The continuous bed stationary phases of these columns consist of porous polymeric materials generated from silica or organic materials such as acrylamide, styrene, acrylate, or methacrylate monomers which result in lower back-pressure than packed particles. The lower back-pressure allows for the use of longer columns and hence greater efficiencies. Several groups have reported on the use of up to 1 m capillary columns (Que and Novotny 2002; Legido-Quigley et al. 2003; Tolstikov et al. 2003; Tanaka et al. 2004) and this technology looks promising.

Plate number and efficiency are also related to particle size (d_p) and column length (L) as shown in Eq. (3). This equation shows that decreasing the particle size increases the theoretical plate number/efficiency (MacNair et al. 1997, 1999; Tolley et al. 2001; Shen et al. 2005). However, Eq. (5) shows again that pressure increases with smaller particle size. Fortunately, new commercial ultra-high pressure liquid chromatography pumps (UPLC) are now available from multiple manufacturers that allow the use of smaller particles in the range of 1−2 μm. These instruments offer substantial resolution enhancements with plate numbers on the order of several hundred thousand and peak capacities in excess of 400 (Wilson et al. 2005). In addition to increased resolution, UPLC also offers higher speed separations as the optimum flow velocity has a significantly broader range which allows for increased flow rates without significant loss of resolution (Wilson et al. 2005). Estimates of up to ninefold increases in flow rates without significant loss of resolution have been suggested (Wilson et al. 2005). It is important to note that ultra-high pressure separations result in increased frictional heating; however this can be reduced by down-scaling the chromatography dimensions with the heating being negligible in columns of less than 1 mm (MacNair et al. 1997).

4 Future Directions and Forward-Looking Thoughts

Although several of the above principles can be used to achieve enhanced chromatographic resolution, the resolution enhancements are still far from that which is needed for very complex metabolomics mixtures. To separate these mixtures, peak capacities of thousands to tens of thousands are necessary. Currently, only multidimensional chromatographic methods offer peak capacities of this magnitude (Mondello et al. 2002; Evans and Jorgenson 2004). Multidimensional chromatography utilizes combinations of two or more separation mechanisms with different selectivity, e. g. ion-exchange and reverse-phase or capillary electrophoresis and reverse-phase LC. These systems offer enhanced resolution due to the utilization of multiple columns with independent chemistries which expands the selectivity of the method. Recall that selectivity improvements can dramatically improve resolution. The maximum peak capacity of a multidimensional system is the product of the two or more individual separation dimensions. For example, a realistic system that has a peak capacity in the first dimension (n_y) of 150 and the peak capacity in the second dimension (n_z) of 50, then the total maximum peak capacity of the multidimensional system is $n_y \times n_z = 150 \times 50 = 7500$. If one considers that an individual metabolome consists of 15,000 metabolites, then one recognizes that this is a considerable increase in comprehensive coverage.

Multidimensional LC-LC separations have been capitalized upon in the area of proteomics and are often referred to as multidimensional protein identification technology (i. e. MUDPIT; Washburn et al. 2001; Wolters et al. 2001); however multidimensional separations have only recently been pursued for metabolomics using GC×GC/time-of-flight (TOF)-MS (Welthagen et al. 2005). Unfortunately, these complex separations will come with increased analysis times, but I believe they will be worth the additional temporal costs.

The above discussion focuses on homogenous multidimensional separations (i. e. LC×LC/MS or GC×GC/MS, but multidimensional LC×GC separations are possible. In fact, the combination of these technologies is commonly referred to as unified chromatography (Chester and Parcher 2001; Chester and Pinkston 2002; Wells et al. 2002, 2003; Luo et al. 2003) and often associated with supercritical fluid chromatography (Chester and Parcher 2001; Chester and Pinkston 2002; Mondello et al. 2002; Wells et al. 2002, 2003; Luo et al. 2003). Although this technology is conceptually exciting, it is still somewhat empirically limited. Another possible LC×GC approach would be to couple HPLC with ion mobility mass (IMS) TOF-MS spectrometry (Verbeck et al. 2002; Guevremont 2004; Liu et al. 2004; Shvartsburg et al. 2005). In this configuration, analytes are ionized as they elute from the HPLC and an electrostatic field propels the analyte ions through a gas field maintained at elevated, atmospheric, or sub ambient pressures. Ions of different size and geometric structure traverse the gas field at different rates dependent upon their charge and collisional cross section therefore allowing separation. The LC-IMS method has been demonstrated for proteomics (Lee et al. 2002; Matz et al. 2002; Liu et al.

2004) and more recently applied to metabolite analyses (Kapron et al. 2005). Extension of this concept to metabolomics will surely occur.

The above text discusses multidimensional chromatographic approaches in an on-line context. However, multidimensional approaches can also be pursued in an off-line, multiplexed, or parallel approach. For example, fractions can be collected off-line using a separate HPLC. The fractions can then be concentrated and reinjected onto an on-line LC/MS system. Alternately, fractions of the same samples could be injected onto a series of parallel systems using different methods (i. e. GC/MS, LC/MS, or various selective modes of each performed with different column selectivities). This is our current approach. For example, samples are fractionated and/or enriched and then the polar and lipophilic fractions are analyzed by GC/MS. In addition methanolic extracts are analyzed for phenolic/saponin content. An interesting concept would be to design a multiplexed system, with multiple chromatographic-mass spectrometry systems operating in an integrated manner. For example, a multiplexed chip system with each chip having a slightly different selectivity and independent mass analyzer could be designed to increase the comprehensive coverage. Such a system with on-line enrichment could also be used to address dynamic range limitations that currently exist for specific compound classes such as phytohormones.

If higher resolution chromatography is obtained, mass analyzers must also be employed with compatible scans speeds to record data for compounds eluting in very short temporal periods. It is expected that LC peak widths of 1–5 s will be routine in the very near future. For accurate quantification, it is commonly accepted that the sampling rate should be sufficient to capture 10 data points across the eluting peak with higher sampling rates being beneficial. Thus, sampling rates should be less than 0.1 s or greater than 10 Hz. This is achievable with current TOF-MS analyzers. It is worth mentioning that quadrupole based mass analyzers, including traps, can approach these speeds; however, TOF mass spectrometers equipped with delayed extraction and ion-reflectrons also offer improved mass accuracy over quadrupoles.

Improvements in the accuracy of the mass analyzer can further enhance metabolite differentiation, elemental composition determination, identification, and allow for the profiling of greater numbers of metabolites. Mass accuracy is directly related to the mass resolution or the ability of the mass analyzer to resolve compounds of different m/z values. Mass resolution is defined in Eq. (6) and is a function of mass (M) divided by the peak width (ΔM) which is most commonly defined at half-height:

$$R_m = \frac{M}{\Delta M} \tag{6}$$

Often, LC/MS is performed with ion-traps or quadrupole mass analyzers that yield mass accuracies in the range of 1.0–0.1 Da. Unfortunately, many metabolites have similar nominal masses which can not be differentiated at this level of mass accuracy. For example, the important natural products genistein and

medicarpin have similar nominal masses of 270, but have different accurate masses of 270.2390 ($C_{15}H_{10}O_5$) and 270.2830 ($C_{16}H_{14}O_4$) respectively due to different chemical compositions. If one could measure their mass with sufficient accuracy, then one could differentiate these compounds in the mass domain even if they could not be physically separated in the chromatographic domain. This mass differentiation can be achieved at a mass resolution ($M/\Delta M$) greater than 6136. Compounds with closer accurate masses such as rutin ($C_{27}H_{30}O_{16}$ = 610.5180) and hesperidin ($C_{28}H_{34}O_{15}$ = 610.5620) would require a higher mass resolution of 13,864 for their differentiation. Mass resolutions on the order of 10,000 can be achieved with modern TOF-MS analyzers, and resolutions in excess of 100,000 with sub-part-per-million mass accuracies (i. e. less than 0.001 at m/z of 1,000 Da) are achievable with Fourier transform ion cyclotron mass spectrometry (FTMS). Newer technologies, such as Thermo Electron Corporation's Orbitraps are currently surfacing that also offer high-resolution solutions. Although high resolution accurate mass measurements have great advantages, this technology is still rather costly.

Interestingly, a significant argument can be made that accurate mass measurements significantly reduce the need for ultra-high resolution separations due to the enhanced separation in the mass domain. However if the chromatography step is omitted or compressed significantly, then ion suppression, competitive ionization, and other matrix affects become increasingly more problematic. I personally believe that both improved chromatographic resolution and accurate mass measurements offer the best solution and that the combination of these techniques will provide greater comprehension and confidence in our ability to profile the metabolome. Further, I also believe that the needed magnitude of enhancements in chromatographic resolution can only be achieved with multidimensional approaches.

References

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci 9:418–425

Birkemeyer C, Kolasa A, Kopka J (2003) Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. J Chromatogr A 993:89–102

Broeckling CD, Huhman DV, Farag MA, Smith JT, May GD, Mendes P, Dixon RA, Sumner LW (2005) Metabolic profiling of Medicago truncatula cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. J Exp Bot 56:323–336

Chester T, Parcher JF (2001) Blurring the Boundaries. Science 291:502–503

Chester T, Pinkston J (2002) Supercritical fluid and unified chromatography. Anal Chem 74:2801–2811

Dixon RA (2001) Phytochemistry in the genomics and post-genomics eras. Phytochemistry 57:145–148

Djordjevic N, Houdiere F, Fowler P (1998) High temperature and temperature programming in capillary HPLC. Biomed Chromatogr 12:153–154

Djordjevic N, Fitzpatrick F, Houdiere F, Lerch G, Rozing G (2000) High temperature and temperature programming in capillary electrochromatography. J Chromatogr A 887:245–252
Djordjevic NM, Fowler PWJ, Houdiere F (1999) High temperature and temperature programming in high-performance liquid chromatography: Instrumental considerations. J Microcolumn Separ 11:403–413
Evans C, Jorgenson J (2004) Multidimensional LC-LC and LC-CE for high-resolution separations of biological molecules. Anal Bioanal Chem 378:1952–1961
Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant fuctional genomics. Nat Biotechnol 18:1142–1161
Guevremont R (2004) High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry. J Chromatogr A 1058:3–19
Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730
Hartman T, Kutchan TM, Strack D (2005) Evolution of metabolic diversity. Phytochemistry 66:1198–1199
Kapron J, Jemal M, Duncan G, Kolakowski B, Purves R (2005) Removal of metabolite interference during liquid chromatography/tandem mass spectrometry using high-field asymmetric waveform ion mobility spectrometry. Rapid Commun Mass Spectrom 19:1979–1983
Lee Y, Hoaglund-Hyzera C, Srebalus Barnes C, Hilderbrand A, Valentine S, Clemmer D (2002) Development of high-throughput liquid chromatography injected ion mobility quadrupole time-of-flight techniques for analysis of complex peptide mixtures. J Chromatogr B Anal Technol Biomed Life Sci 782:343–351
Legido-Quigley C, Marlin N, Melin V, Manz A, Smith N (2003) Advances in capillary electrochromatography and micro-high performance liquid chromatography monolithic columns for separation science. Electrophoresis 24:917–944
Liu X, Plasencia M, Ragg S, Valentine S, Clemmer D (2004) Development of high throughput dispersive LC-ion mobility-TOFMS techniques for analysing the human plasma proteome. Brief Funct Genomic Proteomic 3:177–186
Luo Z, Xiong Y, Parcher J (2003) Chromatography with dynamically created liquid "stationary" phases: methanol and carbon dioxide. Anal Chem 75:3557–3562
MacNair J, Lewis K, Jorgenson J (1997) Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. Anal Chem 69:983–989
MacNair J, Patel K, Jorgenson J (1999) Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0-micron particles. Anal Chem 71:700–708
Matz L, Dion H, Hill H (2002) Evaluation of capillary liquid chromatography-electrospray ionization ion mobility spectrometry with mass spectrometry detection. J Chromatogr A 946:59–68
Mondello L, Lewis AC, Bartle KD (2002) Multidimensional Chromatography. Wiley, Chichester, UK
Oliver DJ, Nikolau B, Wurtele ES (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metab Eng 4:98–106
Patel K, Jerkovich A, Link J, Jorgenson J (2004) In-depth characterization of slurry packed capillary columns with 1.0-microm nonporous particles using reversed-phase isocratic ultrahigh-pressure liquid chromatography. Anal Chem 76:5777–5786
Que A, Novotny M (2002) Separation of neutral saccharide mixtures with capillary electrochromatography using hydrophilic monolithic columns. Anal Chem 74:5184–5191
Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23:131–142
Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13:11–29
Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes M, Willmitzer L, Fernie A, Kopka J (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579:1332–1337

Shen Y, Zhang R, Moore R, Kim J, Metz T, Hixson K, Zhao R, Livesay E, Udseth H, Smith R (2005) Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000–1500 and capabilities in proteomics and metabolomics. Anal Chem 77:3090–3100

Shvartsburg A, Tang K, Smith R (2005) Optimization of the design and operation of FAIMS analyzers. J Am Soc Mass Spectrom 16:2–12

Somerville C, Dangl J (2000) Plant biology in 2010. Science 290:2077–2078

Somerville C, Somerville S (1999) Plant functional genomics. Science 285:380–383

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Takats Z, Wiseman JM, Gologan B, Cooks RG (2004) Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. Science 306:471–473

Tanaka N, Kimura H, Tokuda D, Hosoya K, Ikegami T, Ishizuka N, Minakuchi H, Nakanishi K, Shintani Y, Furuno M, Cabrera K (2004) Simple and comprehensive two-dimensional reversed-phase HPLC using monolithic silica columns. Anal Chem 76:1273–1281

Tolley L, Jorgenson J, Moseley M (2001) Very high pressure gradient LC/MS/MS. Anal Chem 73:2985–2991

Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolitichic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. Anal Chem 75:6737–6740

Trethewey RN (2001) Gene discovery via metabolic profiling. Curr Opin Biotechnol 12:135–138

Trethewey RN, Krotzky AJ, Willmitzer L (1999) Metabolic profiling: a rossetta stone for genomics? Curr Opin Plant Biol 2:83–85

Verbeck G, Ruotolo B, Sawyer H, Gillig K, Russell D (2002) A fundamental introduction to ion mobility mass spectrometry applied to the analysis of biomolecules. J Biomol Tech 13:56–61

Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62:887–900

Washburn M, Wolters D, Yates J (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247

Weckwerth W (2003) Metabolomics in systems biology. Annu Rev Plant Biol 54:669–689

Wells P, Zhou S, Parcher J (2002) Gas-liquid chromatography with a volatile "stationary" liquid phase. Anal Chem 74:2103–2111

Wells P, Zhou S, Parcher J (2003) Unified chromatography with CO2-based binary mobile phases. Anal Chem 75:18A–24A

Welthagen W, Shellie RA, Spranger J, Ristow M, Zimmermannn R, Fiehn O (2005) Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOF) for high resolution metabolomics: biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice. Metabolomics 1:65–73

Wilson I, Nicholson J, Castro-Perez J, Granger J, Johnson K, Smith B, Plumb R (2005) High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. J Proteome Res 4:591–598

Wolters D, Washburn M, Yates J (2001) An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73:5683–5690

I.3 Plant Metabolomics Strategies Based upon Quadrupole Time of Flight Mass Spectrometry (QTOF-MS)

H.A. Verhoeven[1,2], C.H. Ric de Vos[1,2], R.J. Bino[1,2], and R.D. Hall[1,2]

1 Introduction

The growing interest in the use of metabolomics technologies in plant research has come about both due to the broad value of such approaches in almost every field of plant science and also through the improvements in instrumentation and bioinformatics tools which has been realised in recent years. A comprehensive overview of all the various technologies available is beyond the scope of this chapter but the reader is referred to other chapters in this volume or to a number of recent reviews for information on the different approaches (Fiehn 2001, 2002; Sumner et al. 2003; Goodacre et al. 2004). However, MS-based strategies, and in particular in combination with GC or LC separation technology, are proving most popular as these combine very high analytical precision with an equally high detection sensitivity. This enables reliable measurements to be made down to the femtomolar range (Fernie 2003). Furthermore, recent advances in electronics and computing have given rise to the development of yet a new generation of mass spectrometers to supplement the traditional magnetic sector and scanning quadrupole instruments that have been around for several decades now. In this new generation, instruments based on ion traps and time-of-flight (generally referred as TOF) are the most prominent. In particular, the TOF instruments have become popular due to their relatively simple construction and their capacity to be combined with a number of other technologies to enable multi-dimensional analysis. This has resulted in an unprecedented expansion of our metabolomic capabilities. For example, the fast spectral acquisition capacity of TOF instruments has resulted in approximately 1000 components being detected in leaf extracts and an analytical capacity of 1000 samples per month has been achieved (Weckwerth et al. 2001). Such sample numbers and breadth of metabolite detection represent the arrival of true metabolomics research in the true sense of the word. Since that time, our capacity for metabolomic analyses has continued to improve. The high mass accuracy, high resolution, good dynamic range and the large diversity of detectible masses possible with TOF instruments, in association with their intrinsic high sensitivity, are therefore the main reasons behind the many applications, first in the field of proteomics, and now also in metabolomics.

[1] Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands
[2] Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands, e-mail: robert.hall@wur.nl

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

2 The Technology

Time of flight was already introduced in the early 1960s but was quickly replaced by other approaches. This was due to the lack of sufficiently fast electronics needed to process data on a nanosecond scale. Thirty years later in the early 1990s, the development of high megahertz and even gigahertz digital circuits led to the dramatic increase in the application of TOF technology. This, combined with new developments in the area of sample introduction and ionisation of (macro)molecules, has subsequently led to many new applications of (TOF-based) mass spectrometry in the fields of biology and pharmacy.

A TOF instrument serves as the main mass analyser, and its principle is based on ions with different mass/charge ratios having different flight times in a field-free drift zone once they have been accelerated by a very short electric pulse from the electrodes of an accelerator: lighter ions travel faster through the measurement chamber than the heavier ones. A thorough discussion on the physical principles can be found in, for example, Guilhaus (1995). In most TOF instrument designs, ions are detected using Micro Channel Plates (MCP), which, on capturing the ions, generate a cascade of electrons to amplify the signal so that it can be detected by the associated electronics (see Fig. 1 for a schematic representation of the various parts). Several ion recorders have been used with the various designs of hybrid TOF mass spectrometer. The two most widely used are the time-to-digital converter (TDC) and the transient recorder or analogue to digital converter (ADC) (Chernushevich et al. 2001). The type of detector affects both the dynamic range of the signals that can be measured and also the mass accuracy. In a TDC, every individual ion generates a pulse. This pulse is shaped into a digital signal, of which the rising flank is used for timing. The time passed since the start of the ion accelerating pulse and its arrival at the MCP is stored in the memory. This system is very accurate over the entire mass range and is optimally suited for the accurate timing of low ion counts. It is, however, less suitable or even unsuitable for the detection of ions arriving simultaneously at the MCP since these will be recorded as being single events and this will thus lead to an underestimation of the signal. TDCs also suffer from an additional limitation concerning the detection of ions. During the time required to process one pulse, the detector is 'blind' to new incoming pulses. This so-called dead time, not only leads to a further underestimation of the signal, but also it causes a shift in the observed m/z value towards lower values. This can lead to serious deviations from the true accurate mass at high to very high signal intensities. These problems occur to a lesser extent in instruments equipped with an ADC, since these machines can sample the analog output of the MCP at very high frequencies, thus providing multiple data points per observed m/z value. In this way, multiple ions arriving at the same time will lead to a linear increase in peak area. In some designs, TDC and ADC are both used to combine the high mass accuracy of the TDC at low ion rates, with the high dynamic range and accurate m/z value measurements of an ADC at high ion rates.

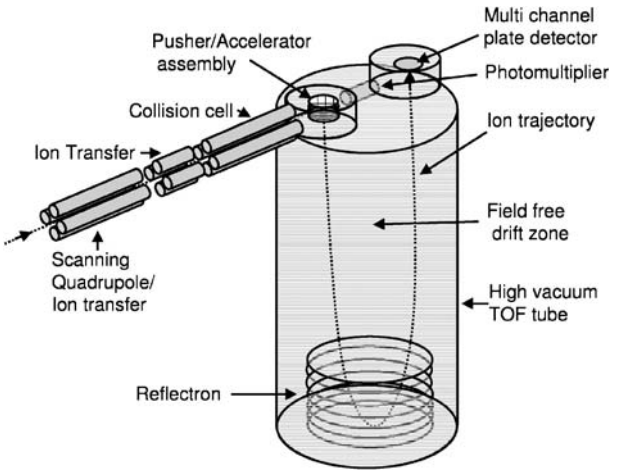


Fig. 1. Schematic diagram showing the main components of a typical quadrupole/time of flight configuration. Ions enter the instrument on the left, and pass through the first quadrupole. This can be operated either in ion transfer mode, which allows all ions to pass, or in selective mode, which is used for precursor scanning and alignment of the different quadrupoles. The ion transfer is a special quadrupole, intended to separate the operating pressures in the different compartments of the quadruple section. In the collision cell, a collision gas can be present in order to induce fragmentation of the incoming ions. When no gas is present molecular ions will be detected. Subsequently, molecular ions and/or charged fragments enter the TOF tube, where they are collected and pushed into the drift zone. During this transition, ions are accelerated in an electric field in the accelerator assembly, consisting of several ion lenses, which determines their kinetic energy. The ions now all follow a trajectory towards the reflectron, consisting of a pile of cylindrical ion lenses at different potentials, which causes them to be repelled towards the detector, the multi channel plate. Here the ions strike the surface of the detector, which finally converts the arrival of every single ion into a measurable electric current. Additional electronics is required to process the electrical signals and the timing between pushing the ions into the drift zone and their arrival at the detector

Regardless of the high mass accuracy, resolution and sensitivity, the application of TOF instruments in structure elucidation is quite limited due to the absence of filtering and scanning capabilities. Consequently, hybrid instruments have since been designed to cope with these shortcomings. These machines include the addition of ion trap(s), quadrupoles or combinations thereof, to the basic TOF analyser. One key example is the now increasingly well-known and widely used QTOF system. These instruments rely on the combination of two or more quadrupoles with a TOF analyser. The first quadrupole (Q1) serves as a mass filter or ion tunnel, depending on the operational mode, with the second quadrupole (Q2) serving as the collision cell for the fragmentation of the ions which have passed through Q1. This fragmentation is achieved using an electric field to accelerate the ions, in combination with a collision gas such as nitrogen or argon. Fragmentation can be controlled by varying the (very low) pressure of the collision gas and/or by varying the collision energy

through altering the acceleration voltage of the cell. Collisions with the gas molecules also result in a cooling of the ions, which incurs that their kinetic energy is transferred. This results in a more homogeneous energy distribution of the individual ions, which in turn improves the mass accuracy capacity of the instrument. The ions and/or ion fragments are subsequently collected in the accelerator part of the TOF instrument where a very short pulse is applied to the electrodes of the chamber to eject the ions. In the case of orthogonal ejection, the differences in kinetic energy in the z-axis will be less than in case of forward ejection. The differences in kinetic energy are also further reduced in the reflectron lens, which repels the ions towards the detector. Here, ions with higher energy will travel further than lower energy ions, thus reducing the difference. A number of variations on this basic design have been created. These include, for example, the modification of the second quadrupole into a linear ion trap with axial ejection through the addition of a number of extra ion lenses (Hager 2002). As a result, new possibilities are created, such as the ability to store specific ions, which can be selectively ejected for complex MS/MS or MSn analyses (Hager 2002). The high mass accuracy can be further improved by using an internal (reference) standard that is sampled at regular intervals throughout the entire analysis period. This reference is then used to correct the instrument calibration on-the-fly (lock mass correction). Such a capacity for continuous (re)calibration is particularly useful, if not essential, in the case of long series of chromatographic runs where excellent, long-term stability of the mass accuracy can be continuously achieved down to ±5 ppm. This is significant as mass accuracy at or below this level allows us to predict the chemical composition of a given ion by using the small known differences in atomic masses of the various atomic elements. In this way, a first prediction can be made about the nature and identity of the molecular component. Combined with other data (retention time, N rule, stable isotope distribution of ^{13}C etc.) this can then enable the list of possible molecular identities to be reduced even further and thus come closer to translating MS output into named metabolites.

Combining the results obtained from several biological samples into a single comparative analysis is an arduous task that requires the precise alignment and matching of peaks representing the same compound over all chromatograms. Due to its relatively robust chromatography and compound separation efficiency, GC-(TOF)-MS of derivatized extracts is at present generally preferred over LC-MS in metabolomic studies (Fiehn et al. 2000; Roessner et al. 2001a,b; Fernie et al. 2004). Nevertheless, GC-MS is less suitable for semi-polar compounds among which are key classes of plant (secondary) metabolites including flavonoids, (glyco-)alkaloids, glucosinolates and saponins. Recent advances in techniques for improving resolution in LC by using capillary electrophoresis (Soga et al. 2002), hydrophilic interaction columns (Tolstikov and Fiehn 2002) and monolithic columns (Tolstikov et al. 2003) demonstrate the high potential which TOF technology has for LC-MS to complement GC-MS in unravelling metabolic profiles.

3 Data Analysis

Data analysis is perhaps the most crucial step in any metabolomics strategy and the importance of bioinformatics tools should not be underestimated. In a standard (ideal) approach, a whole range of standards would be used to assist in identifying through simple linkage, which peaks in an MS output represent which metabolites. However, as the vast majority of metabolites present in complex plant extracts are as yet unknown and are not commercially available, as is especially true for the secondary plant metabolites, this approach is unfeasible at present for a true untargeted metabolomics approach. Another strategy is therefore required which enables the automated and essentially blind direct comparison of large numbers of spectra. Since most datasets are very complicated, dedicated metabolomics software is needed for this purpose. Some of this software is already available but more still needs to be developed and this represents a major task for the next five years.

Data manipulation is essential for reliable metabolomics analyses and special attention has to be paid to aspects such as baseline correction and noise elimination. In addition, in the case of LC-MS, particular attention also needs to be given to reliable correction of local drifts in retention time and accurate mass. Different compositions of eluant can cause significant variation in baseline especially when using steep LC gradients. For the successful correction of such baseline fluctuations, the chromatogram has to contain a region without strong peaks. Digital filtering will enable the elimination of excess noise which would otherwise lead to the generation of erroneous (false) peaks. Some recent software packages are able to deal with a number of these problems in TOF data analysis. Another key element is the need to correct for retention time fluctuations. Unlike capillary gas chromatography which is generally very stable, liquid chromatography often suffers from relatively large, non-linear (localised) fluctuations in retention time. This can be due to small differences in pH, temperature, or the co-elution of components which interact differently with the stationary phase. Consequently, this problem prevents a simple direct comparison of different samples. A number of algorithms have been designed to correct for this phenomenon. One such approach, based on photodiode array type data, uses correlation optimised warping of the chromatograms to achieve alignment of shifted peaks in the chromatograms (Nielsen et al. 1998). For MS data, MetAlign™ software, in contrast, uses specific mass peaks with strong local maxima throughout the chromatogram as 'landmark peaks' with which to correct for chromatographic shifts over the entire series of analyses (Vorst et al. 2005). After correction, unbiased, direct spectral comparisons, based on mass peak intensities are possible and contrasting mass signals can be reliably identified and extracted. Differential chromatograms are produced from which all unchanging peaks have been removed to reveal the true extent of the differences between two (groups of) samples in one or both directions. This dedicated software can automatically handle hundreds of full scan MS datasets obtained by either LC or GC, and is independent of type of mass

spectrometer. Another package, Markerlynx™, which is dedicated to Waters instruments, exploits the high mass accuracy and resolution of the (Q)TOF technology. Here, the distribution of specific ions in a predefined mass window, usually in a 20–50 ppm range, over the chromatogram is analysed and retention shifts are corrected within a predetermined retention time window. In this way, metabolites can be compared over many samples using high mass resolution. Both approaches have their own advantages and limitations. In our lab we more or less routinely use metAlign™ to process LC-QTOF and GC-TOF data. However, in cases where there are insufficient landmark peaks, metAlign™ is unable to perform a thorough alignment. When spectra are noisy or highly complex, the Markerlynx™ approach will likely give misalignments.

How to proceed further with the processing of corrected chromatograms is dependent upon the type of metabolomics analysis required. Within metAlign™ there is a tool to extract user-defined significant differences between two groups of samples based on the Student t-test. In many cases, where large sample numbers are being compared, multivariate analysis will be desirable. For this purpose, software originally developed for the analysis of microarray datasets can be advantageously applied, since metabolomic data share a number of problems similar to those which were encountered with microarray data. We use, for example, the GeneMaths software package (Vorst et al. 2005), which enables rapid statistical data analysis and provides clear graphic outputs of the results in the form of histograms and principle components plots. Other software packages should be equally valuable.

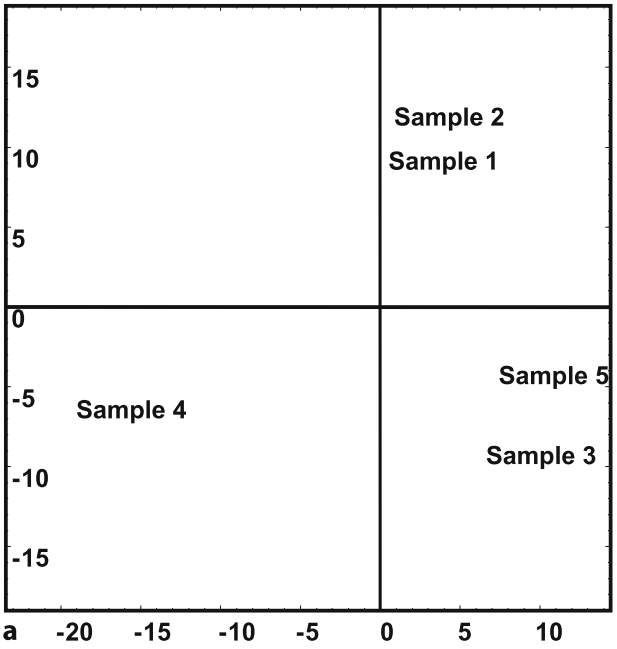
4 Application of QTOF MS-based Plant Metabolomics Analyses

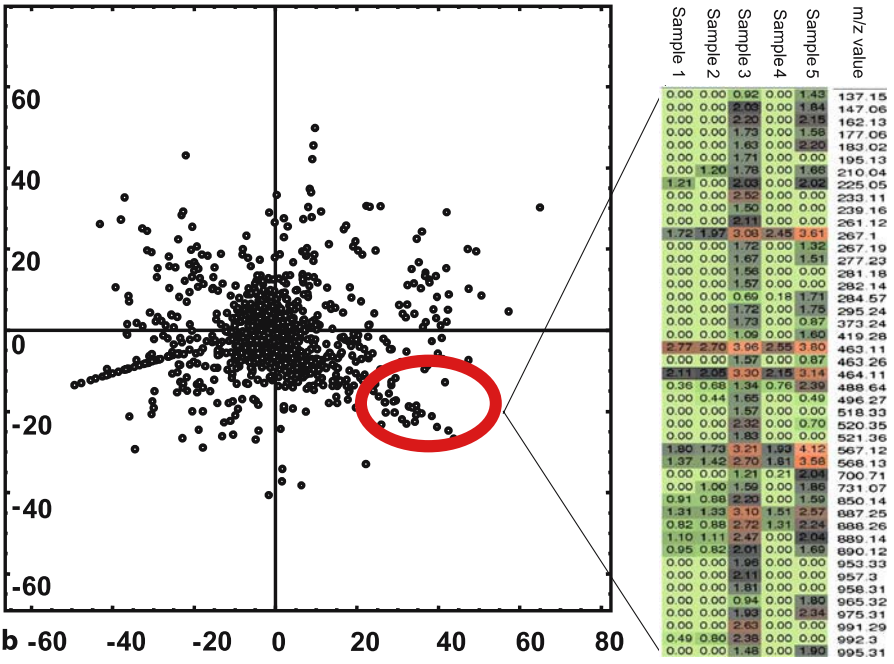
4.1 Rapid Sample Profiling by Direct Flow Injection Analysis (DFI)

In plant metabolomics there is often an initial desire for a rapid pre-screening of the samples. This is especially the case when dealing with large sample numbers where only a limited number of individuals might be expected to be different. One can think here for example, of natural populations, potential mutants or the progeny from a breeding cross (Bino et al. 2005; Hall et al. 2005). Direct Flow Injection can effectively be used to get a rapid, overall impression of the composition of a biological extract. It is an unbiased analysis, and seeks to cover as many metabolites as possible in a single short run. The only selective property is the type of ionisation used, i. e. positive vs negative, ESI (ElectroSpray ionisation) vs APCI (Atmospheric Pressure Chemical Ionisation). The advantage of this approach is time-saving. When using chromatographic separation to prevent excessive component interaction, run times of 30–90 min are usually required (but see as an exception; Jander et al. 2004). In the case of DFI, run times of only 30 s to a few minutes (Goodacre et al. 2003) may be required. For DFI, a few microlitres of extract is introduced by the

autosampler directly into the ion source and all ions with the corresponding charge are then analysed by the MS. In this case, TOF instruments have a clear advantage over scanning quadrupole instruments because their significantly higher resolution allows for the simultaneous detection of many ion species and, because no scanning is required, every individual ion can theoretically be captured. This inevitably results in a very rich mass spectrum which is further complicated by the many interactions which can occur between the different components of the sample during ionization. Furthermore, unstable ions can cause additional extensive ion vapour phase interactions. For these reasons, there was initially considerable scepticism of the potential value of DFI approaches for reliable metabolic analysis and these phenomena are extensively discussed elsewhere (Kebarle 2000; King et al. 2000). However, recent publications have shown that the mass spectrum data obtained in this way is actually highly reproducible and can effectively be used for a fast screening of complex extracts (Aharoni et al. 2002; Goodacre et al. 2002, 2003; Castrillo et al. 2003; Verhoeven et al., in preparation).

Data processing is in many cases the bottleneck for the successful deployment of this technology, and many applications rely on a dedicated approach to data processing. This was clearly demonstrated for example, by the MS analysis of unfractionated plant extracts of *Pharbitis* leaf sap (Goodacre et al. 2003). Correct data processing of the complex mass spectra was found crucial for reliable discrimination between the different physiological treatments used. Experiments performed in our laboratory resulted in similar conclusions. Five commercially available extracts from *Salix* were analysed using DFI in a QTOF MS in positive mode. A single total ion count (TIC) injection peak was observed, and all the masses obtained were combined into a single mass spectrum per sample. The aligned spectra were processed for noise elimination, baseline correction and then centroided to obtain the accurate masses of each m/z peak. These were then aligned in the m/z dimension using exact masses of known metabolites to correct for small fluctuations in exact mass due unavoidable minor (thermal) drift in the TOF tube. Intensities of the m/z peaks were log transformed, and exported to GeneMaths™ for multivariate analysis. Principle Components Analysis (PCA) revealed first (Fig. 2a) that the sample replicates (Samples 1 and 2) cluster close together reflecting the high reproducibility of the extraction and mass profiling techniques. Sample 3 is also clearly similar in overall composition to Sample 5, whereas Sample 4 is clearly distinct from all others. Sample 4 was found to have come from a different supplier. Differences in sample composition were readily detected by selecting the m/z values which were responsible for the separation of the samples in the PCA (Fig. 2b). This example indicates the usefulness of rapid screening for quality control of complex extracts without the need for more dedicated but time-consuming LC separation. In a similar manner Goodacre also used a rapid DFI approach to compare olive oil samples and to test for adulteration (Goodacre et al. 2002).


15
10
5
0
-5
-10
-15
Sample 2
Sample 1
Sample 4
Sample 5
Sample 3
a -20 -15 -10 -5 0 5 10


60
40
20
0
-20
-40
-60
b -60 -40 -20 0 20 40 60 80
Sample 1
Sample 2
Sample 3
Sample4
Sample 5
m/z value

◀ **Fig. 2. a** PCA plot of the entire set of detected mass peaks of 5 *Salix* samples. Samples 1 and 2 were experimental replicates taken from plant extracts of the same origin, but with different batch numbers. Samples 3, 4 and 5 were samples of unknown, but different origin. This figure shows that experimental variation (Samples 1 and 2) is low, Samples 3 and 5 are highly similar with respect to their overall composition while Sample 4 is distinctly different from the rest being placed on the other side of the PCA plot. **b** Detailed PCA of all mass peaks in the Samples 1 to 5. The area responsible for the grouping and positioning of Samples 3 and 5 in the *bottom right quadrant* is highlighted, and the corresponding mass peaks are shown as their logarithmic ratio on the right together with the m/z values of each. *Light grey*: low abundant mass peaks, *dark grey*: highly abundant mass peaks

4.2 QTOF MS Coupled to HPLC

As outlined above, direct infusion methods for a (Q)TOF-MS approach are relatively fast and simple approaches for obtaining metabolic composition fingerprints of multiple samples which can be used to get a preliminary estimation of the extent of similarities and differences between complex extracts. However, ion suppression phenomena may result in decreased detection sensitivity of some compounds, especially of those which ionize relatively poorly (discussed in Kebarle 2000). Moreover, the unavoidable consequences of direct infusion such as matrix-dependent ion suppression, adduct formation and unintended in-source fragmentation, may severely hamper the further detailed interpretation of the origins of the differential mass signals detected. This will thus limit possibilities for the subsequent metabolite identification involved. In addition, with DFI analyses it is, per definition, impossible to discriminate between molecular isomers or between a quasi-molecular ion and an ion having identical mass but which resulted from unintended in-source fragmentation. When such problems arise, or when a more detailed analysis of interesting samples (preselected, e. g. using a DFI approach) is required, LC separation can be used to reduce or solve some of these problems.

Separation of metabolites in complex extracts by liquid chromatography, prior to MS analysis, takes more time but nevertheless has a number of clear advantages. Sensitivity of detection for most compounds will be increased, the formation of adducts at the ionization source will be reduced and the detection of isomeric compounds will be improved. Isomer discrimination is especially important in plant metabolomics as plants are well-known to contain many (secondary) metabolites that may have identical accurate mass but different molecular structures. This is especially true for the large group of flavonoids, within which many compounds have the same elemental composition (and consequently the same accurate mass) whereas the chemical structures are quite distinct, e. g. kaempferol and scutellarein both of which have a neutral accurate mass of 286.04721. Furthermore, when using chromatographic separation, it is also possible to collect additional valuable structural information by applying, e. g. on-line tandem MS and/or by making use of other molecular characteristics such as UV-Vis absorbance and fluorescence which can be

detected on-line prior to the molecules entering the MS. It is this combination of technologies which has made QTOF-based MS analysis a popular choice.

4.2.1 HPLC-PDA-ESI-QTOF-MS

The key to successful, full scale metabolomics analysis is the establishment of a technology platform which generates the maximum amount of reliable information in a single analytical run. For example, the LC-TOF MS based metabolomics system used in our laboratory incorporates a Waters Alliance 2795 HT autoinjector and HPLC pump system fitted with a column oven, a Waters 2996 photodiode array detector (PDA to give absorbance spectra in the 190−700 nm range) and a QTOF Ultima API mass spectrometer with MS/MS capability. In this system, four sets of data are therefore obtained simultaneously: UV/vis spectra, retention time, accurate mass and, when applied, MS/MS fragment information. A lock mass™ spray module is routinely connected to the ESI source in order to correct, on-the-fly, for any small measurement deviations from the exact mass (e. g. Wolff et al. 2001). We routinely use the synthetic peptide leucine enkephalin, which is continuously supplied by a separate low-flow HPLC pump as a reference lock mass in both positive and negative ESI measurement modes. By making combined use of accurate mass, MS/MS fragmentation information and absorbance characteristics, the number of possible elemental compositions and isomers can be narrowed down and essential structural information about a specific metabolite can be derived. For instance, most plant extracts contain multiple (poly)phenolic compounds, among which many isomers exist. Upon MS/MS fragmentation, many isomeric forms can already be distinguished, e. g. quercetin-rhamnoside (m/z 449.1079 in ESI positive mode) provides a fragment of 303.0505 while kaempferol-glucoside (also m/z 449.1079) provides a positively-charged fragment of 287.0556. However, with a QTOF, fragmentation experiments are not always conclusive. For instance, the glycosylated flavonoids kaempferol-3-*O*-glucoside and cyanidin-3-*O*-glucoside have identical mass and elemental composition of $C_{21}H_{20}O_{11}$, and show more or less similar MS/MS fragmentation patterns in ESI-positive mode with the loss of the glucosidic group leaving $C_{15}H_{10}O_6$ as the major fragment. However, in contrast, their UV-Vis absorbance characteristics are markedly different, with only the red-coloured cyanidin-glycoside having significant absorbance at wavelengths between 500 and 520 nm. This additional PDA information is therefore key to rapid isomer discrimination in this case.

A typical chromatogram obtained by reversed phase HPLC separation of a crude plant extract and subsequent on-line detection of eluting compounds by both PDA and QTOF MS is shown in Fig. 3. The observed mass of the metabolite eluting at retention time 23.48 min was 611.1596. Taking into account the uneven mass (indicating an even number of nitrogen atoms) and the isotopic distribution (indicating the absence of sulphur atoms), about 36 different elemental compositions are possible at 5 ppm accuracy (9 at 1 ppm accuracy)

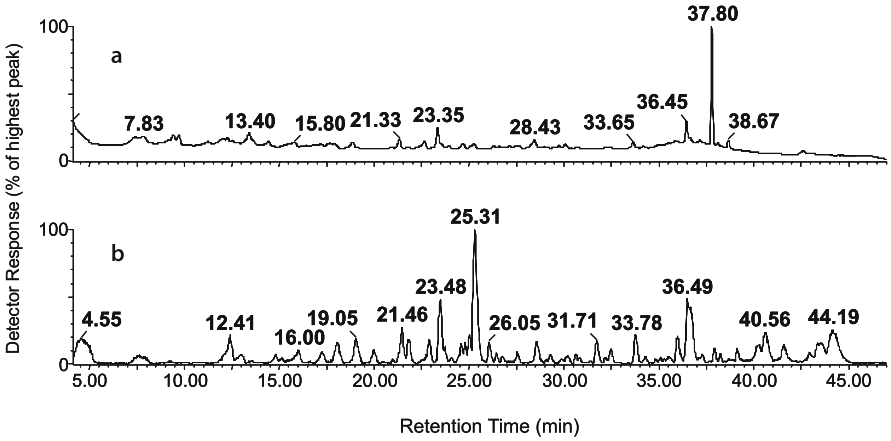


Fig. 3. Typical LC-PDA-QTOF MS chromatograms (base peak intensities) obtained by injection of 5 µl of an aqueous-methanol extract of tomato peel: **a** photodiode array signal (240–600 nm) **b** QTOF-ESI$^+$-MS signal (m/z 100–1500). Indicated are retention times of the most intense peaks

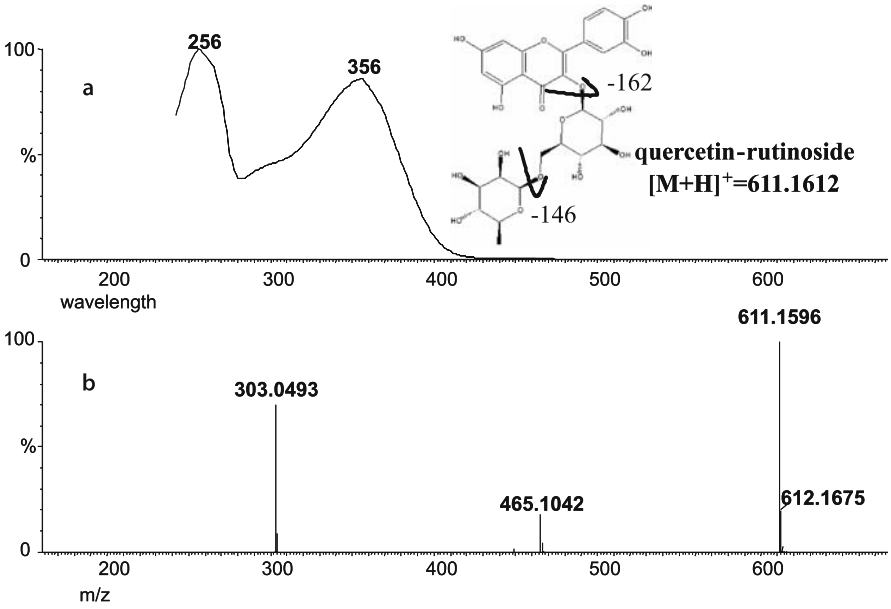


Fig. 4. a Absorbance spectrum of chromatographic peak at retention time 23.48 min. **b** QTOF-ESI-MS/MS spectrum of chromatographic peak at retention time 23.48 min. Observed accurate mass of parent ion $[M+H]^+$ = 611.1596 corresponds to an elemental composition of $C_{27}H_{31}O_{21}$ (−2.6 ppm) and its fragments obtained correspond to $C_{21}H_{21}O_{16}$ (+1.9 ppm) and $C_{15}H_{11}O_7$ (−3.9 ppm)

with parameter settings of $C \leq 75, H \leq 100, O \leq 75, N \leq 10$ and $P \leq 4$. Subsequent on-line LC-MS/MS fragmentation experiments (Fig. 4b) showed neutral losses of 146 and 162, and accurate mass fragments ($[M+H]^+$) of 465.1042 and 303.0493. The accurate mass and MS/MS fragmentation pattern correspond to a metabolite having an elemental composition of $C_{27}H_{30}O_{21}$, e. g. a diglycosylated anthraquinone, an (iso)flavonoid or a benzoyl-benzoic acid. The PDA spectrum of the same chromatographic peak (Fig. 4a) showed two absorbance maxima at around 255 and 360 nm, indicative of a flavonol-type flavonoid with at least two hydroxyl groups on the B ring (Markham 1989). The combination of accurate mass, MS/MS fragments and UV/vis absorbance spectrum indicates that the most likely candidate is quercetin-3-*O*-rutinoside. This is also supported by the knowledge that this flavonoid has been reported to be the major flavonol in tomato fruits (Muir et al. 2001; Le Gall et al. 2003). Subsequent comparison with an authentic standard indeed revealed identical chromatography, accurate mass, MS/MS fragmentation pattern as well as absorbance spectrum thus confirming peak identity.

4.2.2 Metabolomics to Characterize Tomato Mutants

The LC-PDA-QTOF-MS-based platform approach has been shown to be an effective, reproducible and sensitive method for non-targeted metabolomic profiling. Sample preparation, chromatographic system and accurate mass measurements have been optimized in order to screen hundreds of extracts in an unsupervised stable manner. After unbiased alignment of the chromatograms, the data are imported into multivariate analyses software to elucidate the biological variables underlying the data structure (Vorst et al. 2005). As an example of the power of non-targeted metabolomics approaches using LC-QTOF MS, we recently reported on the effect of a single mutation on the metabolic profile of ripe fruits of tomato (Bino et al. 2005). In tomato, several natural photomorphogenic mutants are known and these have been the subject of detailed physiological investigations. One mutant, carrying one of the *high pigment* (*hp-1*, *hp*-1^w, *hp*-2, *hp*-2^j, and *hp*-2^{dg}) mutations, is characterized by its exaggerated light responsiveness. Generally, these mutants have higher pigmentation levels in their hypocotyls, leaves and fruit in comparison to their semi-isogenic, wild-type counterparts (Levin et al. 2003; van Tuinen et al. 2005). The more intense colour of the fruits is a clear indication that these mutants accumulate more all-trans lycopene in their ripe fruits. However, by using a metabolomics approach it became clear that the metabolic perturbations in these fruit were much more extensive than just involving lycopene (Bino et al. 2005). The *hp*-2^{dg} mutant and wild-type tomato plants (cv. Manapal) were grown simultaneously under controlled environmental conditions and fruit samples were pooled per plant. Different, complementary metabolic profiling techniques, including GC-MS and LC-MS, were applied to measure as many

compounds as possible in ripe fruits. For non targeted profiling of non-volatile (semi-polar) compounds, aqueous methanol extracts were prepared and subjected to reversed phase HPLC using both PDA (240–600 nm) and QTOF-MS (m/z 100–1500; ESI positive and negative modes). A lock mass spray (sampled every 10 s) was used to enable accurate mass measurements. Raw data were processed by the metAlign™-software and mass traces were extracted (6168 in negative mode and 5401 in positive mode with a ratio of > 3) and aligned across all samples. Pair-wise comparisons (Student t-test) could then be used to determine statistically significant differences between *hp*-2^{dg} and wild-type fruit extracts. Differential chromatograms were produced from the original LC-MS software, and from these (Fig. 5) it was evident that the *hp*-2^{dg} mutation resulted in a significant increase in many compounds (246 mass signals in negative mode and 137 in positive mode were > twofold higher) and a decrease in only a small number of other compounds (57 mass signals being a factor 2 or more lower in negative mode and 5 mass signals twofold lower in positive mode). The metabolites corresponding to the differential masses were identified using accurate mass, MS/MS fragmentation experiments and absorbance spectra (PDA) information. In this way, it was possible to identify a number of phenolic compounds, flavonoids and alkaloids that were significantly increased in the *hp*-2^{dg} mutant (Bino et al. 2005) pointing clearly to a pleiotropic

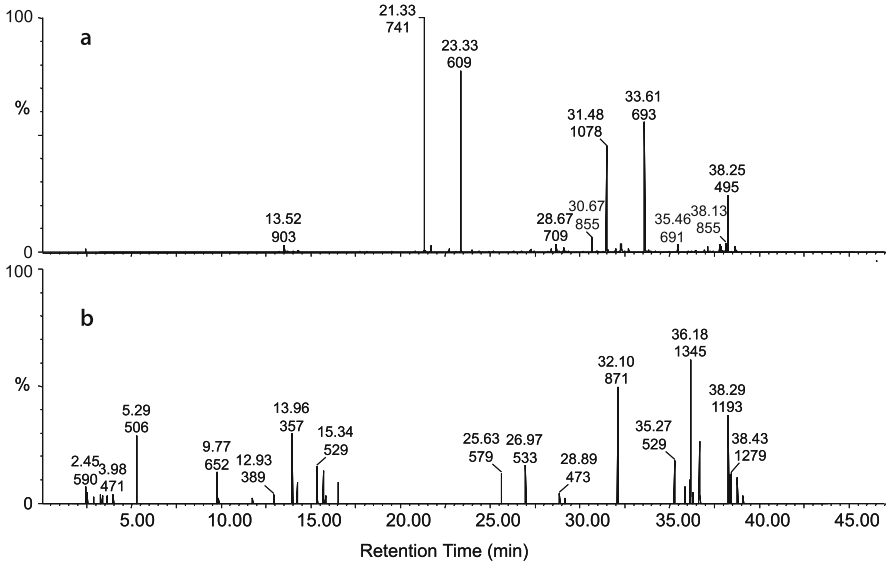


Fig. 5a,b. Metalign™-processed LC-QTOF MS chromatograms (recorded in ESI-negative mode) showing metabolites that are significantly different (Student t-test, $p < 0.01$; $n = 5$) between *hp*-2^{dg} and wild-type tomato fruits: **a** metabolites at least twofold higher in *hp*-2^{dg} than in wild-type; **b** metabolites at least twofold higher in wild-type than in *hp*-2^{dg}. Retention times and nominal masses of metabolites are indicated. 100% scale of y-axis (TIC) is 25,000 in **a** and 500 in **b**

effect of photomorphogenic mutations on tomato fruit metabolism which was much greater than was initially visible.

5 Conclusions and Future Prospects

The examples presented in this chapter clearly underline the versatility of hybrid TOF mass spectrometers, and their capabilities with regard to metabolic profiling, structure elucidation and compound identification, using accurate mass determinations and MS/MS fragmentation. The high sensitivity and mass resolution allows the rapid screening of complex plant extracts by DFI, suitable for semi quantitative high throughput (pre)screening. More detailed analysis is possible when MS detection is preceded by applying separation technologies such as LC. Data processing and efficient data handling are becoming more and more the bottleneck in the process, especially when high throughput screening is required and the currently available bioinformatics tools are inadequate. Another bottleneck is the low number of available reference compounds needed for definitive identification of differentially accumulating components. Key developments for the near future will therefore have to be made in these areas if true plant metabolomics strategies are to become routine. With better software and more easily mined databases we will be best equipped for the identification of the large numbers of the highly chemically diverse components typically present in complex plant extracts.

References

Aharoni A, de Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, Bino RJ, Goodenowe DB (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclon Mass Spectrometry. OMICS 6:217–234

Bino RJ, de Vos CHR, Lieberman M, Hall RD, Bovy A, Jonker HH, Tikunov Y, Lommen A, Moco S, Levin I (2005) The light-hyperresponsive high pigment-2^{dg} mutation of tomato: alterations in the fruit metabolome. New Phtyologist 166:427–438

Castrillo JI, Hayes A, Mohammed S, Gaskell SJ, Oliver SG (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. Phytochemistry 62:929–937

Chernushevich IV, Loboda AV, Thomson BA (2001) An introduction to quadrupole-time-of-flight mass spectrometry. J Mass Spectrom 36:849–865

Fernie AR (2003) Metabolome characterization in plant system analysis. Funct Plant Biol 30:111–120

Fernie AR, Trethewey RW, Krotzky AJ, Willmitzer L (2004) Metabolic profiling: from diagnostics to systems biology. Nature Rev Mol Cell Biol 5:763–769

Fiehn O (2001) Combining genomics, metabolome analysis and biochemical modelling to understand metabolic networks. Comp Funct Genom 2:155–168

Fiehn O (2002) Metabolomics-the link between genotypes and phenotypes. Plant Mol Biol 48:115–171

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolic profiling for plant functional genomics. Nat Biotechnol 18:1157–1161

Goodacre R, Vaidyanathan S, Bianchi G, Kell DB (2002) Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. Analyst 127:1457–1462

Goodacre R, York EV, Heald JK, Scott IM (2003) Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. Phytochemistry 62:859–863

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolomics data. Trends Biotechnol 22:245–252

Guilhaus M (1995) Principles and instrumentation in Time-of-flight mass spectrometry. J Mass Spectrom 30:1519–1532

Hager JW (2002) A new linear ion trap mass spectrometer. Rapid Commun Mass Spectrom 16:512–526

Hall RD, de Vos CHR, Verhoeven HA, Bino RJ (2005) Metabolomics for the assessment of functional diversity and quality traits in plants. In: Harrigan G, Vaidyanathan S, Goodacre R (eds) Metabolic profiling. Kluwer Acad Publ, Dordrecht, Netherlands pp 31–44

Jander G, Norris SR, Joshi V, Fraga M, Rugg A, Yu S, Li L, Last RL (2004) Application of a high-throughput HPLC-MS/MS assay to Arabidopsis mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. Plant J 39:465–475

Kebarle P (2000) A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. J Mass Spectrom 35:804–817

King R, Bonfiglio R, Fernandez-Metzler C, Miller-Stein C, Olah T (2000) Mechanistic investigation of ionization supression in electrospray ionization. J A Soc Mass Spectrom 11:942–950

LeGall G, DuPont MS, Mellon FA, Davis AL, Collins GJ, Verhoeyen ME, Colquhoun IJ (2003) Characterization and content of flavonoid glycosides in genetically-modified tomato (*Lycopersicon esculentum*) fruits. J Agric Food Chem 51:2438–2446

Levin I, Frankel P, Gilboa N, Tanny S, Lalazar A (2003) The tomato dark green mutation is a novel allele of the tomato homolog of the *DEFOLIATED 1* gene. TAG 106:454–460

Markham KR (1989) Flavones, flavonols and their glycosides. In: Dey PM, Harborne JB (eds) Methods in plant biochemistry, vol 1. Academic Press, San Diego, USA, pp 197–235

Muir S, Collins GJ, Robinson S, Hughes S, Bovy A, de Vos CHR, van Tunen AJ, Verhoeyen ME (2001) Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonoids. Nat Biotechnol 19:470–474

Nielsen N-PV, Carstensen JM, Smedsgaard J (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A 805:17–35

Roessner U, Willmitzer L, Fernie AR (2001a) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. Plant Physiol 127:746–764

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A (2001b) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13:11–29

Soga T, Ueno Y, Naraoka H, Matsuda K, Tomita M, Nishioka T (2002) Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. Anal Chem 74:6224–6229

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combining of hydrophilic interaction chromatography and electrospray ion trap spectrometry. Anal Biochem 301:298–307

Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolithic silica-based capillary reversed phase liquid chromatography / electrospray mass spectrometry for plant metabolomics. Anal Chem 75:6737–6740

Van Tuinen A, de Vos CHR, Hall RD, van der Plas LHW, Bowler C, Bino RJ (2005) Use of metabolomics for identification of tomato genotypes with enhanced nutritional value derived from natural light-hyperresponsive mutants. In: Jaiwal PK (ed) Improving the nutritional and therapeutic qualities of plants. (Plant Metabolic Engineering & Molecular pharming.) SciTech Publishers, Raleigh, USA (in press)

Vorst OF, de Vos CHR, Lommen A, Staps RV, Visser RGF, Bino RJ, Hall RD (2005) A non directed approach to the differential analysis of multiple LC/MS-derived metabolic profiles. Metabolomics 1:169–180

Weckwerth W, Tolstikov V, Fiehn O (2001) Metabolomic characterization of transgenic potato plants using GC/TOF and LC/MS analysis reveals silent metabolic phenotypes. Abstract: Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics (1–2)

Wolff JC, Eckers C, Sage AB, Giles K, Bateman R (2001) Accurate mass liquid chromatography / mass spectrometry on quadrupole orthogonal acceleration time-of flight mass analyzers using switching between separate sample and reference sprays. 2 Applications using the dual-electrospray ion source. Anal Chem 73:2605–2612

I.4 Capillary HPLC

T. Ikegami[1], E. Fukusaki[2], and N. Tanaka[1]

1 Introduction

Among many techniques employed for the separation and identification of metabolites, HPLC (high performance liquid chromatography)-MS (mass spectrometry) is most widely applicable to metabolomics, although the chromatographic efficiency is generally lower than that of the other separation techniques, GC (gas chromatography)-MS or CE (capillary electrophoresis)-MS. Recently, significant improvement was made to increase the separation capability of HPLC, which will help the analysis of complex metabolite samples. In the field of metabolomics, because of the importance of separation and detection of thousands of small molecules, micro HPLC techniques will become a common method of separation in the near future (Tomita and Nishioka 2003). In this article, the use of long capillary columns that give high separation efficiencies in micro HPLC system, and multidimensional HPLC that can provide even higher peak capacity will be described. Special attention will be paid to the examples of high efficiency HPLC separations made possible by monolithic silica columns composed of network type silica skeletons.

2 Monolithic Silica Columns for Micro HPLC

Micro HPLC systems with a monolithic silica capillary column possess the following advantages:

1. Small consumption of stationary and mobile phases
2. High detection sensitivity for a certain amount of samples
3. High speed separation with low pressure drop
4. The possible use of a long column with $1 \sim 2$ m that can provide around $100{,}000 \sim 200{,}000$ theoretical plates

along with some disadvantages:

[1] Department of Polymer Science and Engineering, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto, 606–8585, Japan, e-mail: ikegami@kit.ac.jp, nobuo@kit.ac.jp
[2] Department of Biotechnology, Graduate School of Engineering, Osaka Univ, 2–1 Yamadaoka, Suita, 565–0871, Japan, e-mail: fukusaki@bio.eng.osaka-u.ac.jp

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

Table 1. Column sizes, flow rates, linear velocities, and degrees of sample dilution

Column type	Inner diameter [mm (μm)]	Column volume[a] [μl]	Flow rate [μl/min]	t_0^a [s/10 cm]	Solvent linear velocity [mm/s]	Relative degree of dilution[b]
Conventional	4.6	1660	1000	70	1.4	2100
Semi-micro	2.0	314	200	66	1.5	400
Micro	1.0	78	50	66	1.5	100
	0.5 (500)	20	12.5	66	1.5	25
Micro-capillary	0.3 (300)	7.1	5	59	1.7	9
	0.2 (200)	3.1	2	66	1.5	4
	0.1 (100)	0.78	0.5	66	1.5	1
	0.05 (50)	0.20	0.12	69	1.5	0.25
	0.025 (25)	0.05	0.03	69	1.5	0.06

[a] Column lengths were 10 cm, total porosity was estimated as 0.70
[b] Column of id 100 μm is taken as a standard

1. Smaller sample capacities of a monolithic silica column than particle-packed columns
2. Necessity of skill and knowledge to operate a capillary HPLC system to obtain high separation efficiency, and insufficient supply of good columns and instruments for capillary HPLC

Particle-packed capillary columns have been employed for separations of analytes with or without the assistance of electroosmotic flow. It is possible to pack silica particles into a fused silica capillary equipped with a frit, but it is difficult to produce high efficiency and long-lasting columns using $1 \sim 2$ μm particles (Novotony 1988; Knox and Grant 1991; Schmeer et al. 1995).

Recently, monolithic silica capillary columns have been reported to show higher separation efficiencies than particle packed columns (Ishizuka et al. 2000; Tanaka et al. 2000). They consist of network silica skeletons that can be prepared in capillaries by a sol-gel method. Monolithic silica columns of 4.6 mm ID (inner diameter), 0.2 mm ID, 0.1 mm ID and 0.05 mm ID are commercially available at present. Column sizes and flow rates to be employed are listed in Table 1.

2.1 Characteristics of Monolithic Silica Columns

Here, the features of monolithic silica capillary columns and the optimization of separation conditions will be described. The use of monolithic silica columns consisting of network silica skeletons and through-pores for micro HPLC was reported recently (Minakuchi et al. 1996; Tanaka et al. 2001). Monolithic silica capillary columns were reported to provide better separation efficiencies than particle-packed columns, and the use of these columns for proteomics and

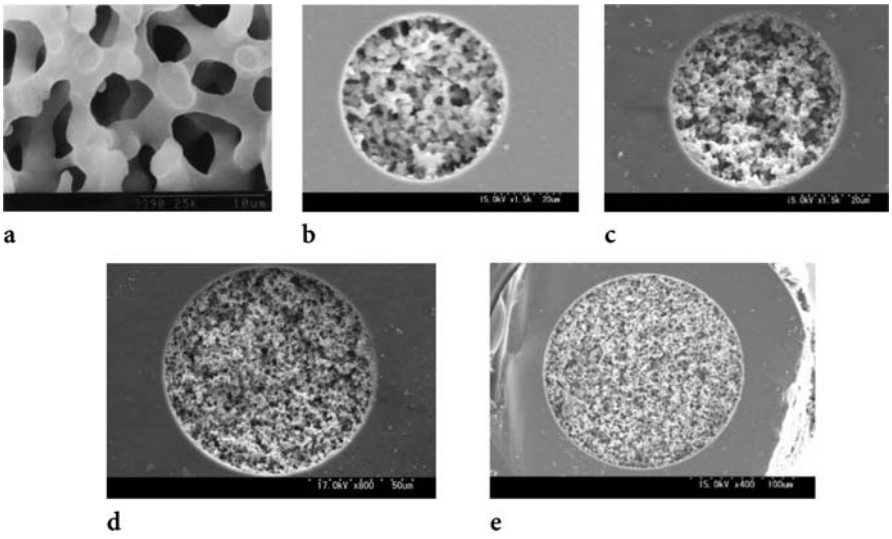


Fig. 1. Scanning electron microscope images of monolithic silica prepared from sol-gel methods: **a** monolithic silica prepared in a test tube; **b,c** monolithic silica prepared in 50 μm ID fused silica capillary; **d** monolithic silica prepared in 100 μm ID fused silica capillary; **e** monolithic silica prepared in 200 μm ID fused silica capillary tube

metabolomics seems to be attractive (Cabrera 2004). Monolithic silica columns are prepared by acid-catalyzed hydrolytic polymerization of alkoxysilanes in the presence of water-soluble polymers such as poly(ethylene glycol) (Tanaka et al. 2001; Cabrera 2004). Figure 1a shows a scanning electron microscope (SEM) image of monolithic silica prepared in a test tube, while Fig. 1b–e shows SEM images of monolithic silica columns prepared in fused silica capillaries with 50 ~ 200 μm internal diameter (Motokawa et al. 2002).

Currently available monolithic capillary columns include organic polymer columns (Svec 2004) and chemically modified silica columns, and they have the following features. Monolithic polymer columns generally show higher permeability than particle-packed columns, and high efficiency for the separation of macromolecules (Svec et al. 2000). In the case of monolithic silica capillary columns, silica skeletons are covalently bonded to capillary walls. Thus, frits are not necessary to hold the skeletons in a column, and column length can be varied in the range of 5 ~ 200 cm after preparation. Generally, the silica skeleton sizes are in the range of 1 ~ 2 μm. As shown in Fig. 1a, monolithic columns have 3 ~ 10 times bigger (through-pore size/skeleton size) ratio, 1 ~ 3, than particle-packed columns with (through-pore size/particle size) ratio, 0.25 ~ 0.4. Monolithic silica columns produce similar separation efficiencies to particle-packed columns at much lower pressure drop. At the same pressure drop, monolithic columns can provide higher separation efficiencies than particle-packed columns. Moreover, due to the small silica skeleton sizes,

relatively high separation efficiencies can be expected at higher linear velocities (Minakuchi et al. 1997, 1998). In terms of separation impedance, total performance of columns (E), monolithic silica capillary columns can produce higher separation efficiencies, nearly 10 times greater than that of a particle-packed column (Motokawa et al. 2002). Separation impedance is given by Eq. (1) where N, ΔP, t_0 and η stand for number of theoretical plates, column back pressure, the elution time of an unretained solute, and viscosity of mobile phase, respectively (Bristow and Knox 1977):

$$E = t_0 \Delta P / N^2 \eta = (\Delta P / N)(t_0 / N)(1/\eta) \quad (1)$$

Figure 2 shows chromatograms produced by monolithic silica capillary columns of 25 ~ 130 cm lengths modified with C18 stationary phase (Ishizuka et al. 2002).

The dilution factors for analytes are proportional to the internal diameters of columns squared assuming that band broadening (peak width) and resolutions are similar for various HPLC systems. Sample concentrations after the separation are higher in micro columns with smaller diameters, and the higher sample concentrations can lead to higher detection sensitivity (Table 1). Because lower flow rates can lead to higher ionization efficiencies and higher detection sensitivity in LC-ESI (electrospray ionization) MS system, development of high

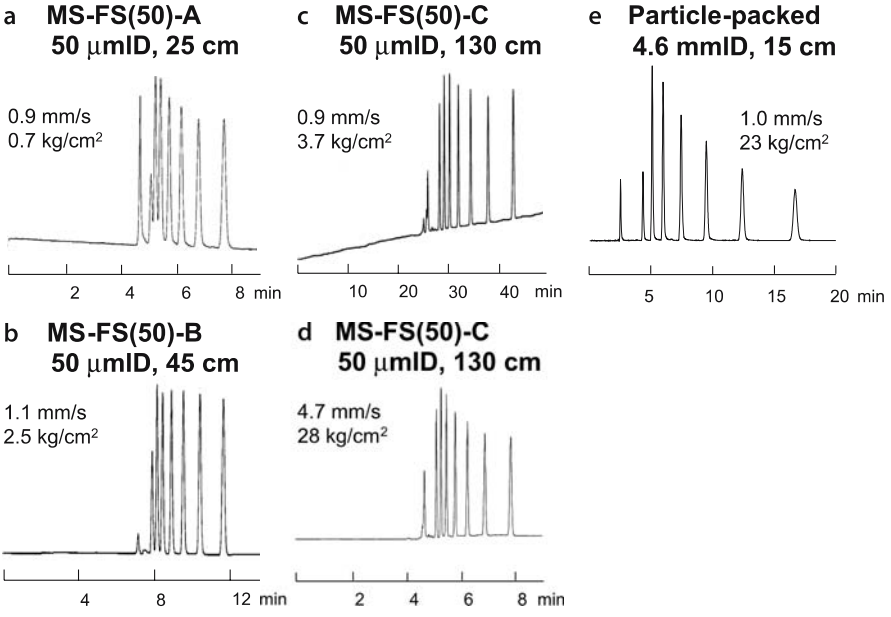


Fig. 2. Chromatograms obtained for alkylbenzenes ($C_6H_5(CH_2)_nH$, n = 0 – 6) by: **a–d** C18 monolithic silica capillary columns; **e** particle packed column (5 mm silica-C18 particles, Mightysil RP18)

efficiency micro HPLC system is an important issue for metabolomics studies (Schmidt et al. 2003).

2.2 Column Efficiencies and the Optimization of Separation Conditions

The number of theoretical plate N is a measure of the quality of a column and elution conditions, and is given by Eq. (2) from the retention time of a peak (t_R) and peak width at half height ($t_{w1/2} = 2.35\sigma$, σ being the standard deviation of a Gaussian peak). Resolution R_s is given by Eq. (4), that includes N, α (Eq. (5), selectivity, the ratio of retention factors of two adjacent peaks), and k (Eq. (3), a retention factor, distribution coefficient of a solute between stationary and mobile phases, i. e. the ratio of times (t_R–t_0) to t_0, the former stands for time the solute exists in mobile phase, and the latter stands for time the solute exists in stationary phase). For convenient separation and detection, the k values should be in a range of 2 ~ 5:

$$N = (t_R/\sigma)^2 = 5.54(t_R/t_{w1/2})^2 = 16(t_R/t_w)^2 \tag{2}$$

$$k = (t_R - t_0)/t_0 \tag{3}$$

$$R_s = (N^{1/2}/4)[(\alpha - 1)/\alpha][k/(1 + k)] \tag{4}$$

$$\alpha = k_2/k_1 \tag{5}$$

$$\Delta P = \phi\eta u L/d_p^2 (u = L/t_0) \tag{6}$$

ΔP is proportional to η, u (linear velocity of the mobile phase), and L (column length) while it is inversely proportional to d_p^2, where d_p stands for diameter of particle. Thus, a column packed with particles of small diameter leads to high separation efficiency, (greater N) at the expense of high column backpressure. Due to the drawback, an approach to get high efficiencies by reducing diameter of particles has a limit: since the pressure limit of a pump system is around 300 ~ 400 bar with a normal operational pressure 100 ~ 200 bar, the limit in particle sizes is in a range of 1 ~ 3 μm. The flow resistance parameter ϕ in Eq. (6), is usually ca. 2000 for particle-packed columns, while ϕ values reach to 200 ~ 400 in the case of monolithic silica columns (Giddings 1965; Bristow and Knox 1977).

A solute band is broadened when it travels outside a column due to parabolic flow profile in a tube as well as due to slow diffusion in the stagnant mobile phase existing in an injector, a detector, or connection tubing. Especially, for solutes of small retention factors which elutes in early part of a chromatogram, sample injection into a capillary column of 1 ~ 5% of column volume has significant influence on band spreading, mainly caused by sample diffusion at orifice in an injector or by dead volume in all connection parts (Ikegami at al. 2004). The split-flow injection technique is practical and useful for micro HPLC with monolithic silica capillary columns in order to avoid the peak spreading during

injection (Taniguchi and Murata 2002). Moreover, the use of weak eluents for sample injection is also effective to increase the separation efficiency: in the case of reversed-phase HPLC, sample solution can be prepared with water-rich solvent (Ikegami at al. 2004).

3 Applications of Monolithic Silica Columns to Metabolomics

Figure 3 shows chromatograms of leaf extracts of *Arabidopsis thaliana* by LC-ESI-MS using 30 ∼ 90 cm monolithic silica capillary columns modified with C18 stationary phase under gradient conditions, from aqueous ammonium acetate buffer (pH 5.5) to acetonitrile, MeCN (Tolstikov et al. 2003). A shallow gradient (large t_G, gradient time) with a long column has lead to better separation. The results indicate that improvement of separation by the use of the longer columns caused the reduction of ion suppression effect by introducing the solute bands separately into ES ionization interface. In the case of Fig. 3, the peak capacity provided by the long monolithic silica column is still not enough for complete separation, but it shows a feasible approach of using longer monolithic silica capillary columns to achieve higher separation efficiency avoiding ion suppression effect in the LC-ESI-MS system. This approach will result in longer separation time, but the amount and quality of information after the analysis of metabolites would be better than conventional LC-MS systems using particle packed columns. Connected monolithic columns (conventional size) in series showed good separation of polyprenol homologues (Bamba et al. 2004).

Mass spectrometry would often be used in metabolomics research due to its superiority in both quantification and qualification. However, mass spectrometry has a serious drawback named 'ionization suppression'. Ionization suppression is a phenomenon that presence of impurity at ionization might cause a serious impairment in qualitative accuracy (King et al. 2000; Müller et al. 2002). Coelution in chromatography might cause ionization suppression. Even the technology of capillary monolithic chromatography might not provide a perfect time separation that is one of the ideal solutions against ionization suppression. Recently, stable isotope dilution technology tends to be used as a practical tool to reduce an ionization suppression negative effect. A stable isotope dilution method employs isotopologues as an internal standard that would be separated not by chromatography but by mass spectrometry to provide an accurate comparative quantification. This principle is used in the proteomics research tool 'isotope coded affinity tags (ICAT)' (Han et al. 2001). A metabolic profiling of sulfur metabolite using ^{34}S was reported (Mougous et al. 2002). ^{13}C and ^{15}N stable isotope labeling techniques could be available in some case. In addition, post sampling stable isotope labeling would be also applicable, although D-labeling may face some difficulties (Zhang et al. 2001; Fukusaki et al. 2005). In future a combination of monolithic capillary

chromatography and stable isotope diluted comparative quantification would be one of the de facto standard methods in metabolomics.

4 Two-Dimensional HPLC

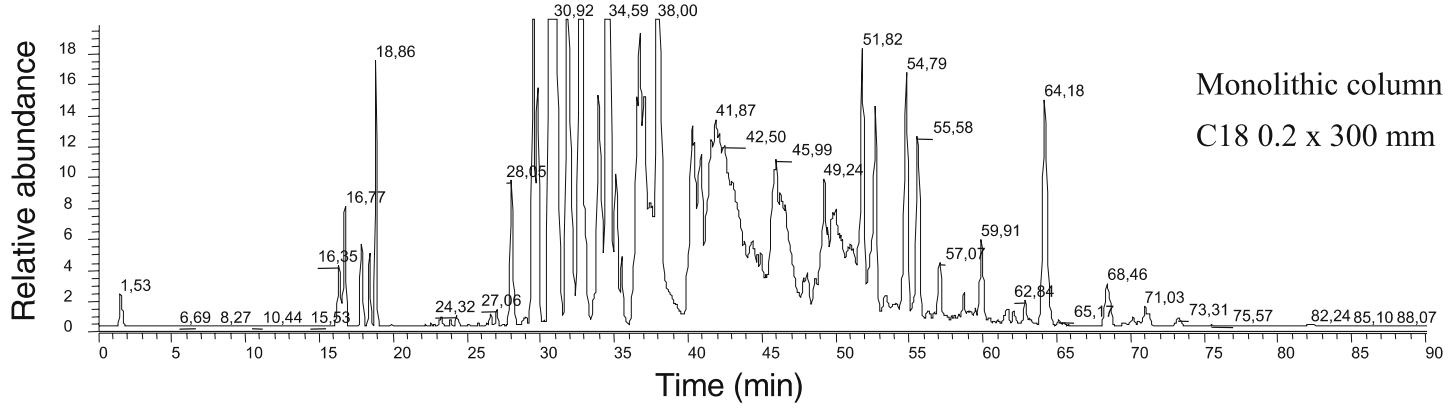
Peak capacity (PC) given by Eq. (7) indicates the separation ability regarding how many solutes can be potentially separated by a chromatographic system. Retention times of the first solute and the last solute are given as t_1 and t_R respectively in Eq. (7). Separation methods such as ultrahigh-pressure liquid chromatography (UHPLC) and supercritical fluid chromatography (SFC) can produce a PC of ca. 300/h (Shen and Lee 1998; MacNair et al. 1999), while a conventional HPLC system gives a PC of 100 ∼ 200/h. In order to achieve far larger PC using conventional HPLC systems, multidimensional separation systems were shown to be effective. When two chromatographic systems with PC_x and PC_y are combined to form a two-dimensional (2D) chromatography system, PC for the total system can be theoretically estimated as a product of two PC values as Eq. (8) (Giddings 1991):

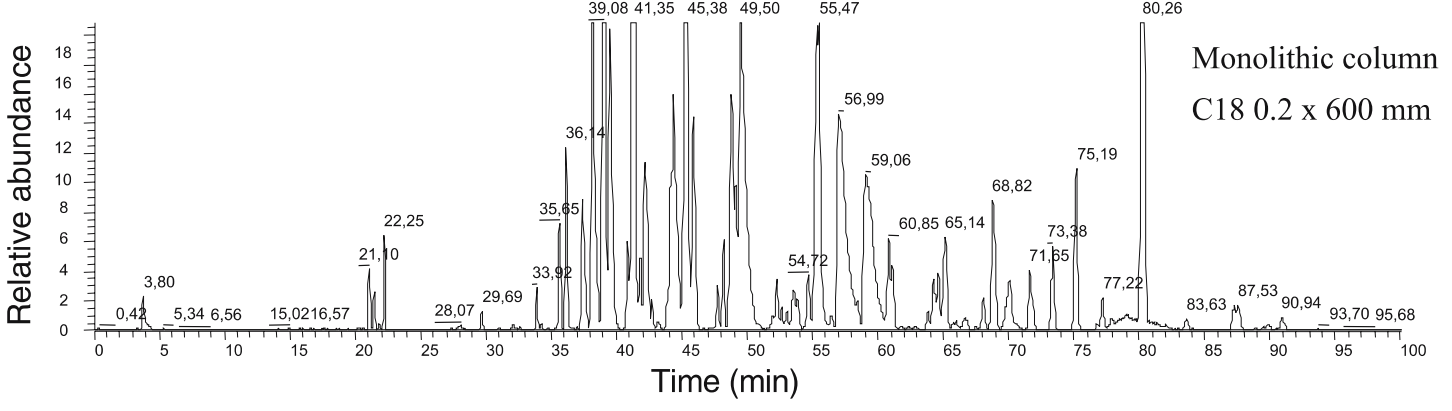
$$PC = 1 + (N^{1/2}/4)\ln(t_R/t_1) \tag{7}$$

$$PC_{2D} = PC_x \times PC_y \tag{8}$$

In comprehensive 2D-HPLC separations every fraction obtained from 1st-D separation is to be separated in 2nd-D HPLC, while the next fraction is eluted from 1st-D. Therefore the 2nd-D column should ideally be eluted at very high speed to meet the rate of fractionation at the 1st-D separation. The 2nd-D column should possess low-pressure drop and reasonable efficiency at high flow rate. In addition to high efficiency and high permeability, the 1st-D and 2nd-D columns must possess adequate difference in selectivity to effect 2D separations. Ideally the 1st-D and 2nd-D should have orthogonal selectivity or different separation mechanisms (Bushey and Jorgenson 1990; Köhne and Welsch 1999; Wagner et al. 2002; Venkataramani and Zelechonok 2003). Ion-exchange mode and reversed-phase mode, or size-exclusion mode and reversed-phase mode have often been combined to effect 2D separations for peptide mixtures in proteomics. Because a particle-packed column cannot be operated at adequately high flow rate, various approaches were taken in the past: (i) small columns were employed at 1st-D compared to 2nd-D, (ii) the first column was eluted slowly or intermittently, or (iii) two or more sets of chromatographs were used at the 2nd-D. Even with these methods, however, truly "two-dimensional" HPLC is hard to achieve due to the mixing of separation modes.

Figure 4 shows a scheme of 2D-HPLC and its working principle, in which the outlet tubing of the 1st-D column was connected to a loop of the 2nd-D injector to couple particle the packed 1st-D column and 2nd-D monolithic silica column run at higher linear velocity (Tanaka et al. 2004). In this case, the


Relative abundance
Time (min)
Monolithic column
C18 0.2 x 300 mm
1,53
6,69
8,27
10,44
15,53
16,35
16,77
18,86
24,32
27,06
28,05
30,92
34,59
38,00
41,87
42,50
45,99
49,24
51,82
54,79
55,58
57,07
59,91
62,84
64,18
65,17
68,46
71,03
73,31
75,57
82,24
85,10
88,07


Relative abundance
Time (min)
Monolithic column
C18 0.2 x 600 mm
0,42
3,80
5,34
6,56
15,02
16,57
21,10
22,25
28,07
29,69
33,92
35,65
36,14
39,08
41,35
45,38
49,50
54,72
55,47
56,99
59,06
60,85
65,14
68,82
71,65
73,38
75,19
77,22
80,26
83,63
87,53
90,94
93,70
95,68

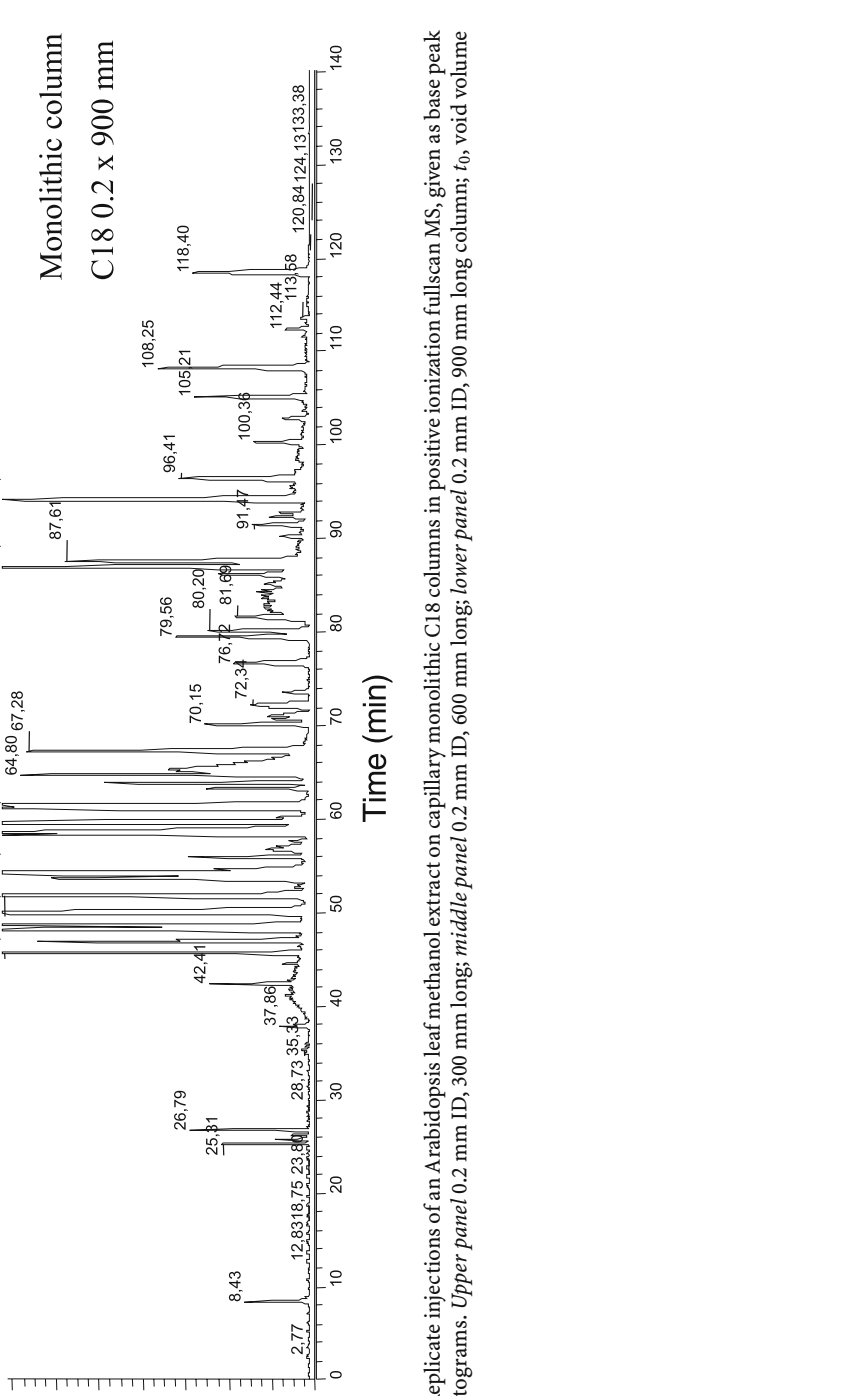


Fig. 3. Replicate injections of an Arabidopsis leaf methanol extract on capillary monolithic C18 columns in positive ionization fullscan MS, given as base peak chromatograms. *Upper panel* 0.2 mm ID, 300 mm long; *middle panel* 0.2 mm ID, 600 mm long; *lower panel* 0.2 mm ID, 900 mm long column; t_0, void volume

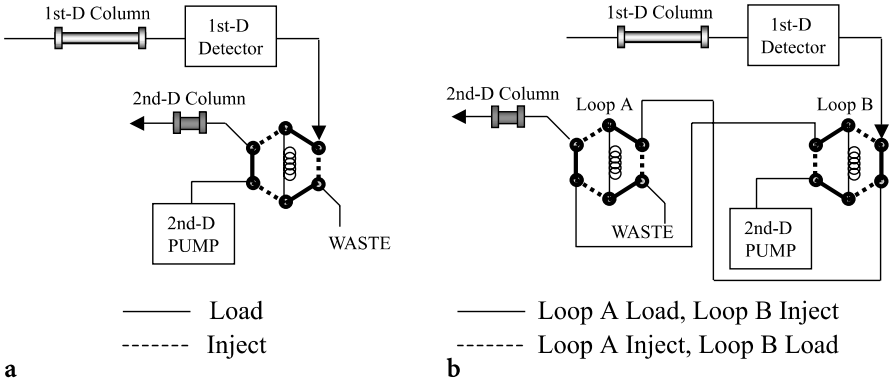


Fig. 4. a Tubing connection at 2nd-D injector of simple 2D-HPLC. **b** Tubing connection of two six-port valves used as 2nd-D injector

fraction from the 1st-D column is loaded and temporarily kept in a loop of the 2nd-D injector that results in mixing of separated peaks, but the flow rates of two HPLC systems can be controlled independently. The 2nd-D separation can be carried out at very high flow rate (for example, 10 ml/min for a 4.6 mm ID column) throughout the separation. The simplest 2D-HPLC in Fig. 4a produced PC = 1000 in reversed phase mode. When two six-port valves or a ten-port valve is used at the 2nd-D HPLC in Fig. 4b, all fractions can be subjected to the separation at the 2nd-D column to provide a comprehensive 2D-HPLC system resulting in so-called group separation, solutes of similar structural features appear as a group. Because of fast flow rate in the 2nd-D separation using a 4.6 mm ID column, the 2D-HPLC system consumed a lot of mobile phase solvent. In order to reduce the consumption of mobile phases, the sufficiently fast, simple 2D-HPLC using capillary columns has been examined (Kimura et al. 2004). The use of capillary column at 2nd-D leads to less solvent consumption and better MS detectability compared to a larger-sized column. Figure 5a shows a 2D chromatogram for the tryptic digest of BSA (Bovine serum albumin) obtained from total ion monitoring by ESI-TOF (Time of flight)-MS. From the 1st-D (2.1 mm ID, 5.0 cm long), 18 fractions were injected at 2-min intervals into the 2nd-D reversed-phase system (4.6 mm ID, 2.5 cm long), generating 18 chromatograms that were used to produce a 2D chromatogram. Figure 5b shows a 2D chromatogram obtained for the separation of tryptic digest of BSA using a capillary column (100 μm ID, 10 cm long) in the 2nd D separation. The number of spots distinguishable in vertical direction in Fig. 5b was greater than that in Fig. 5a. This is due to the higher column efficiency and longer gradient time in 2nd-D, along with greater MS detection sensitivity based on nearly optimum flow rate (3 μL/min) on the capillary column, the greater amount of sample introduced to the 2nd-D column because of the longer fractionation interval, and the smaller extent of dilution due to the use of small diameter column (Kimura et al. 2004).

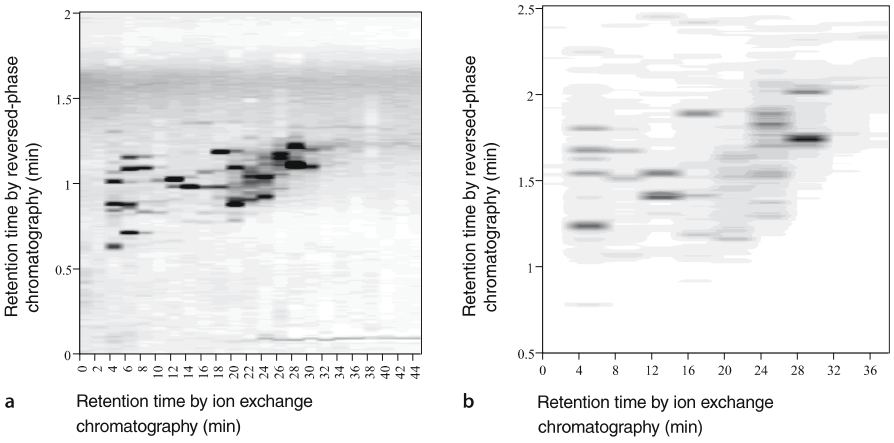


Fig. 5. Two-dimensional separation of tryptic digest of BSA in simple 2D-HPLC, 1st-D; MCI CQK-31S column (2.1 mm ID, 50 mm long), flow rate; 50 µl/min: **a** 2nd-D; monolithic silica-C18 column (4.6 mm ID, 25 mm long), flow rate; 5.0 ml/min; **b** 2nd D; C18 monolithic column (0.1 mm ID, 100 mm long), flow rate in a capillary column; 3.0 µl/min with a split flow/injection; linear velocity in the column; 7.7 mm/s. ESI-TOF-MS detection, total ion chromatogram for a mass range 400–2000

5 Combination of Reversed-Phase HPLC and Other Separation Modes

Since many compounds of similar properties are to be separated in proteomics, 2D-HPLC hyphenated to an MS system has been employed combining ion exchange mode and reversed phase mode, or size-exclusion mode and reversed-phase mode. In the case of metabolomics, combination of several different separation modes is preferable to separate a variety of substances. Reversed-phase mode is most often employed in HPLC, where chemically bonded stationary phases (C8, C18, and C30, etc.) have advantages in rapid equilibration with mobile phase, high separation efficiency, and high reproducibility in gradient.

Recently, hydrophilic LC (HILIC LC) (Alpert et al. 1994; Yoshida 1997) was shown to be effective for the separation of metabolomes utilizing the interaction between solutes and hydrophilic functional groups on the stationary phases. The selectivities of HILIC columns are similar to those of a conventional silica column, but HILIC columns have advantages over silica columns in the recovery of samples, and compatibility with mobile phases used in reversed-phase mode. Figure 6 shows a comparison of elution patterns of HILIC mode and reversed-phase mode in the separation of an extract from *Arabidopsis thaliana* (Tolstikov and Fiehn 2002). Since the solvent type for HILIC and reversed-phase mode are common, it is possible to combine the two separation modes to form multidimensional HPLC, although

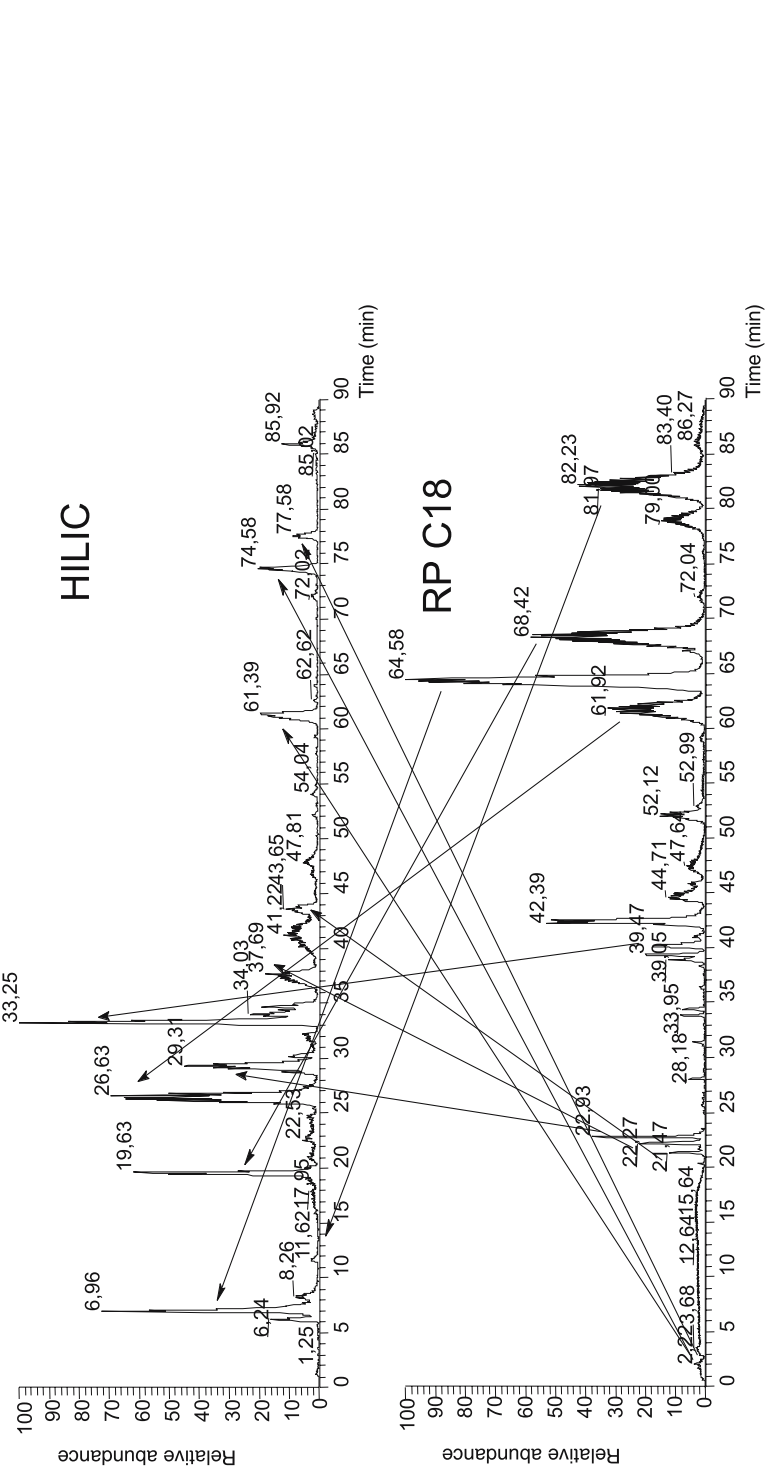


Fig. 6. Comparison of chromatograms of an *Arabidopsis thaliana* leaf methanol extract, obtained by HILIC-LC mode (*top panel*) and reversed-phase mode (*bottom panel*): Conditions (*top panel*) TSK Gel Amide 80, 4.6 mm ID, 150 mm long, gradient elution from MeCN to ammonium acetate buffer (6.5 mmol/l, pH 5.5), MeCN content (%) (time, min) 100 → 100(5) → 90(8) → 60(75) → 0(80), (*bottom panel*) C18 column, 4.6 mm ID, 150 mm long, gradient elution from ammonium acetate buffer (6.5 mmol/l, pH 5.5) to MeCN, MeCN content (%) (time, min) 0→ 0(15) → 95(40) → 100(60) → 100(80)

the compositions of mobile phases that controls the retention order are total opposites to each other. Capillary columns for HILIC LC are under development.

6 Outlook

Routine use of micro HPLC will need development of several important constituents; the reproducible preparation of high performance columns, small-volume pumps and gradient systems, and improvement of an injection system. Subjects to be studied are the development of high performance monolithic silica columns for variety of separation modes, multidimensional microLC systems, and optimization of an interface between LC and MS instruments. Large peak capacities realized by highly efficient microHPLC systems or multidimensional HPLC will greatly contribute to metabolomics studies when coupled with MS instruments and stable isotope dilution methodology.

References

Alpert AJ, Shukla M, Shukla AK, Zieske LR, Yuen SW, Ferguson MAJ, Mehlert A, Pauly M, Orlando R (1994) Hydrophilic-interaction chromatography of complex carbohydrates. J Chromatogr A 676:191–202

Bamba T, Fukusaki E, Nakazawa Y, Kobayashi A (2004) Rapid and high-resolution analysis of geometric polyprenol homologues by connected octadecylsilylated monolithic silica columns in high-performance liquid chromatography. J Sep Sci 27:293–296

Bristow PA, Knox JH (1977) Standardization of test conditions for high performance liquid chromatography columns. Chromatographia 10:279–289

Bushey MM, Jorgenson JW (1990) Automated instrumentation for comprehensive two-dimensional high-performance liquid chromatography of proteins. Anal Chem 62:161–167

Cabrera K (2004) Application of silica-based monolithic HPLC columns. J Sep Sci 27:843–852

Fukusaki E, Harada K, Bamba T, Kobayashi A (2005) An isotope effect on the comparative quantification of flavonoids by means of methylation-based stable isotope dilution coupled with capillary liquid chromatograph/mass spectrometry. J Biosci Bioeng 99:75–77

Giddings JC (1965) Dynamics of chromatography, part 1. Principles and theory. Dekker, New York

Giddings JC (1991) Unified separation science. Wiley-Interscience, New York, pp 126–128

Han DK, Eng J, Zhou H, Aebersold R (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol 19:946–951

Ikegami T, Dicks E, Kobayashi H, Morisaka H, Tokuda D, Cabrera K, Hosoya H, Tanaka N (2004) How to utilize the true performance of monolithic silica columns. J Sep Sci 27:1292–1302

Ishizuka N, Minakuchi H, Nakanishi K, Soga N, Nagayama H, Hosoya K, Tanaka N (2000) Performance of a monolithic silica column in a capillary under pressure-driven and electrodriven conditions. Anal Chem 72:1275–1280

Ishizuka N, Kobayashi H, Minakuchi H, Nakanishi K, Hirao K, Hosoya K, Ikegami T, Tanaka N (2002) Monolithic silica columns for high-efficiency separations by high-performance liquid chromatography. J Chromatogr A 960:85–96

Kimura H, Tanigawa T, Morisaka H, Ikegami T, Hosoya K, Ishizuka N, Minakuchi H, Nakanishi K, Ueda M, Cabrera K, Tanaka N (2004) Simple 2D-HPLC using a monolithic silica column for peptide separation. J Sep Sci 27:897–904

King R, Bonfiglio R, Fernandez-Metzler C, Miller-Stein C, Olah T (2000) Mechanistic investigation of ionization suppression in electrospray ionization. J Am Soc Mass Spectrom 11:942–950

Knox JH, Grant IH (1991) Electrochromatography in packed tubes using 1.5 to 50 μm silica gels and ODS bonded silica gels. Chromatographia 32:317–328

Köhne AP, Welsch T (1999) Coupling of a microbore column with a column packed with non-porous particles for fast comprehensive two-dimensional high-performance liquid chromatography. J Chromatogr A 845:463–469

MacNair JE, Patel KD, Jorgenson JW (1999) Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0 mm particles. Anal Chem 71:700–708

Minakuchi H, Nakanishi K, Soga N, Ishizuka N, Tanaka N (1996) Octadecylsilylated porous silica rods as separation media for reversed-phase liquid chromatography. Anal Chem 68:3498–3501

Minakuchi H, Nakanishi K, Soga N, Ishizuka N, Tanaka N (1997) Effect of skeleton size on the performance of octadecylsilylated continuous porous silica columns in reversed-phase liquid chromatography. J Chromatogr A 762:135–146

Minakuchi H, Nakanishi K, Soga N, Ishizuka N, Tanaka N (1998) Effect of domain size on the performance of octadecylsilylated continuous porous silica columns in reversed-phase liquid chromatography. J Chromatogr A 797:121–131

Motokawa M, Kobayashi H, Ishizuka N, Minakuchi H, Nakanishi K, Jinnai H, Hosoya K, Ikegami T, Tanaka N (2002) Monolithic silica columns with various skeleton sizes and through-pore sizes for capillary liquid chromatography. J Chromatogr A 961:53–63

Mougous JD, Leavell MD, Senaratne RH, Leigh CD, Williams SJ, Riley LW, Leary JA, Bertozzi CR (2002) Discovery of sulfated metabolites in mycobacteria with a genetic and mass spectrometric approach. Proc Natl Acad Sci USA 99:17037–17042

Müller C, Schäfer P, Störtzel M, Vogt S, Weinmann W (2002) Ion suppression effects in liquid chromatography-electrospray-ionisation transport-region collision induced dissociation mass spectrometry with different serum extraction methods for systematic toxicological analysis with mass spectra libraries. J Chromatogr B 773:47–52

Novotny M (1988) Recent advances in microcolumn liquid chromatography. Anal Chem 60:500A–510A

Schmeer K, Behnke B, Bayer E (1995) Capillary electrochromatography – electrospray mass spectrometry: a microanalysis technique. Anal Chem 67:3656–3658

Schmidt A, Karas M, Dülcks T (2003) Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: when does ESI turn into nano-ESI. J Am Soc Mass Spectr 14:492–500

Shen Y, Lee ML (1998) General equation for peak capacity in column chromatography. Anal Chem 70:3853–3856

Svec F (2004) Preparation and HPLC applications of rigid macroporous organic polymer monoliths. J Sep Sci 27:747–766

Svec F, Peters EC, Sýkora D, Yu C, Fréchet JMJ (2000) Monolithic stationary phases for capillary electrochromatography based on synthetic polymers: designs and applications. J High Resolut Chromatogr 23:3–18

Tanaka N, Nagayama H, Kobayashi H, Ikegami T, Hosoya K, Ishizuka N, Minakuchi H, Nakanishi K, Cabrera K, Lubda D (2000) Monolithic silica columns for HPLC, micro-HPLC, and CEC. J High Resolut Chromatogr 23:111–116

Tanaka N, Kobayashi H, Nakanishi K, Minakuchi H, Ishizuka N (2001) Monolithic LC columns. Anal Chem 73:420A–429A

Tanaka N, Kimura H, Tokuda D, Hosoya K, Ikegami T, Ishizuka N, Minakuchi H, Nakanishi K, Shintani Y, Furuno M, Cabrera K (2004) Simple and comprehensive two-dimensional reversed-phase HPLC using monolithic silica columns. Anal Chem 76:1273–1281

Taniguchi H, Murata Y (2002) The newest protocol of proteomics 9, Capillary HPLC. Cell Tech 21:1332–1343

Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. Anal Biochem 301:298–307

Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. Anal Chem 75:6737–6740

Tomita M, Nishioka T (2003) Frontier of metabolomics. Springer, Berlin Heidelberg New York

Venkatramani CJ, Zelechonok Y (2003) An automated orthogonal two-dimensional liquid chromatograph. Anal Chem 75:3484–3494

Wagner K, Miliotis T, Marko-Varga G, Bischoff R, Unger KK (2002) An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. Anal Chem 74:809–820

Yoshida T (1997) Peptide separation in normal phase liquid chromatography. Anal Chem 69:3038–3043

Zhang R, Sioma CS, Wang S, Regnier FE (2001) Fractionation of isotopically labeled peptides in quantitative proteomics. Anal Chem 73:5142–5149

I.5 Capillary HPLC Coupled to Electrospray Ionization Quadrupole Time-of-flight Mass Spectrometry

S. Clemens, C. Böttcher, M. Franz, E. Willscher,
E. v. Roepenack-Lahaye, and D. Scheel[1]

1 Introduction

Metabolite profiling in the pre-metabolomics era of the early 1970s to the late 1990s as well as the pioneering metabolomics projects since the late 1990s have been predominantly GC-MS based. GC-MS techniques are robust and well-established. Many primary metabolites (e. g. organic acids, sugars, amino acids, sugar alcohols) can easily be derivatized and are therefore amenable to GC-MS analysis. Also, spectral databases and deconvolution algorithms are available, which help extracting meaningful information. Early on, however, it was obvious that no single analytical technique would be sufficient to achieve comprehensive coverage of the metabolome (Sumner et al. 2003). As stated from the beginning and reiterated since, the chemical diversity of metabolites makes it virtually impossible to detect all compound classes in one "catch" (Goodacre et al. 2004; Dunn et al. 2005). That is why already the first reports describing GC-MS-based metabolomics platforms emphasized the need to develop complementing LC-MS platforms (Roessner et al. 2000). LC-MS covers in principle a much wider mass range and should allow one to target many compound classes not detectable by GC-MS. Furthermore, there is usually no need for derivatization and LC-MS offers superior options to elucidate unknown metabolites structurally. Particular fractions can easily be collected for NMR analysis and metabolites/molecular ions can be further analyzed by tandem-MS or even MS^n. Hampering the adoption of LC-MS approaches for metabolomics, however, was the fact that LC-MS has only rather recently (i. e. in the 1990s) developed into a routine technology (Niessen 1999a).

One might argue that the need for LC-MS-based profiling is even more pressing in plant science. A highly rich and diverse secondary metabolism is a hallmark of plant biology. Lacking the ability to avoid or to retreat from unfavorable conditions or potential foes, plants have evolved an enormous metabolic plasticity, which allows them to respond dynamically to environmental changes through the synthesis and/or degradation of particular compounds. This is complemented by the accumulation of various pre-formed defenses against microbial attack and other threats (Dixon 2001). Furthermore, many so-called secondary metabolites also apparently play major roles in primary developmental processes and as signaling molecules. Flavonoids

[1] Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle/Saale, Germany, e-mail: sclemens@ipb-halle.de

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

and their biosynthesis, for instance, have long been investigated because of their role in flower pigmentation, UV protection, or pathogen defense (Winkel-Shirley 2001). More recent work demonstrated that flavonoids negatively regulate auxin transport and are required for pollen germination (Taylor and Grotewold 2005).

A large fraction of plant secondary metabolites has been classically analyzed by LC techniques, predominantly through separation on reversed phase material. Thus, it is a straightforward concept to combine this with state-of-the-art mass spectrometry in order to develop powerful metabolomics platforms that cover important compound classes such as phenylpropanoids or alkaloids. A look at *Arabidopsis thaliana*, the most important plant model species, can illustrate the need for and the potential of LC-MS profiling. Because *A. thaliana* has no history of use as a medicinal plant, it initially did not attract the attention of too many natural product chemists. As a consequence, few secondary metabolites were identified 10 years ago. In the course of the genome sequencing, however, it became increasingly clear, that *A. thaliana* should produce thousands of different compounds. The *Arabidopsis* genome encodes a myriad of proteins likely to be involved in secondary metabolism (d'Auria and Gershenzon 2005). There are more than 270 cytochrome P450 genes, more than 100 glycosyl transferase genes, about 50 glutathione *S*-transferase genes, to name a few. For most of the encoded enzymes we do not know substrates or products.

The first major challenge for metabolomics is the huge chemical diversity of the metabolome. The second lies in the fact that – as indicated above for *Arabidopsis thaliana* – most of the metabolites in any given higher eukaryote are unknown. Current estimates are in the range of 4000–20,000 metabolites for a given species (Fernie et al. 2004). Unlike for proteins, genome sequences do not allow one to deduce the structure of the metabolites. Instead, the structure has to be elucidated because for only a very minor portion of the metabolites are standards available. Thus, the future success of metabolomics will also be determined by the ability to identify reliably metabolites and to establish the metabolomes of the important model species. Again, this is a particularly daunting task for plants and filamentous fungi, organisms that synthesize huge numbers of secondary metabolites, many of which might only be synthesized in certain cell types or at particular developmental stages. LC-MS, especially in the combination of quadrupole and time-of-flight analysis in modern hybrid instruments, holds the promise to meet this challenge as well. Structural information can in principal be obtained in three different ways: (i) by determining the elemental composition through the accurate mass, (ii) by exploiting the information provided by in-source fragmentation, and (iii) by performing targeted CID-MS (collision-induced dissociation). In contrast, GC-MS-based profiling faces severe limitations when it comes to de novo identification of unknown compounds (Fiehn 2002). Molecular ions are rarely detected because most analytes are derivatized and molecules are fragmented by the electron impact ionization.

We will in the following discuss the principles of capillary LC-MS-based profiling, describe the current state and present new data from our own laboratory on the optimization and the potential of capillary LC coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (CapLC-ESI-QTOF-MS) (von Roepenack-Lahaye et al. 2004).

2 Extraction, Chromatography and Mass Spectrometry

When optimizing extraction and chromatographic separation of low molecular weight compounds, there are various considerations which are commonplace in analytical chemistry (Niessen 1999b) and which will therefore only be touched upon very briefly. The extraction of biological material with aqueous methanol has so far been the most widely used option for GC-MS as well as LC-MS metabolite profiling schemes (Roessner et al. 2000; Fiehn et al. 2000; Tolstikov et al. 2003). For the sake of stability of compounds and reproducibility of the analysis, cold extraction is preferred in most cases. Obviously, the choice of solvent greatly influences scope and range of the profiling. We tested, for instance, acetonitrile-water and methanol-water mixtures for the extraction of *Arabidopsis thaliana* seeds and counted simply the number of mass signals with a signal to noise ratio > 5 by analyzing subsets of the resulting LC-MS chromatograms with MetAlign (www.metalign.nl) (see below). We detected 1680 mass peaks in an 80% methanolic seed extract and 1771 signals in a 50% methanolic seed extract. Of these signals, 973 were found in both extracts. Utilization of acetonitrile-water gave comparable results: upon extraction with 80% and 50% acetonitrile, 2070 and 1771 mass peaks, respectively, were detected and 1029 were found in both extracts. Only 532 mass peaks were detected in all extracts.

There are several classical analytical options to enrich selectively certain compound classes by modifying the extraction. Solid-phase extraction can be used to remove problematic compounds such as lipids or to concentrate others that are of interest but give low signal intensity. These different options and their effect on the metabolome coverage of LC-MS approaches have not been systematically evaluated yet. A way of selectively targeting specific classes of molecules is derivatization. This can permit analysis of compounds with inadequate stability and results in better chromatographic behaviour as well as enhanced signal intensity. Also, derivatization has been proposed as a means to make the ionization of diverse analytes more uniform by adding a particular chemical group (Halket et al. 2005).

A major obstacle in the development of LC-MS was the general incompatibility of flow rates between LC and MS, i. e. the need to introduce a column effluent of about 1 mL/min into a high vacuum (Niessen 1999a). One solution to this problem was to reduce the flow rate by miniaturization of the LC column, a second to split the column effluent so that only a fraction reaches the

mass spectrometer. Often these two options are combined. In capillary liquid chromatography the flow rate is reduced to meet the optimum flow rate range characteristic for many ESI interfaces. Splitting occurs – if at all – prior to chromatography between the pump and the column. Chromatography is performed at low flow rates of 2–20 µL/min (Abian et al. 1999). Column diameters are typically between 80 and 800 µm. In principle, MS is a mass flow sensitive detection because the response is proportional to the actual number of molecules reaching the detector. However, at a constant flow rate under atmospheric pressure ionization conditions, MS acts as a concentration sensitive detector, i. e. the signal is proportional to the analyte concentration in the eluent (Niessen 1999a). The smaller diameter of a capillary column as compared to a regular 4.6-mm analytical column combined with a lower flow rate allows the use of much smaller sample volumes and lower sample concentrations. Furthermore, depending on the design of the ESI interface a reduced flow-rate can result in higher sensitivity due to the enhanced ionization yield of the smaller primary droplet formation (Wilm and Mann 1994). Thus, since the mid 1990s there has been a trend towards miniaturization of the LC (Abian et al. 1999), although the better sensitivity – i. e. lower concentration detection limits – is partly offset by the need to reduce the injection volume and by the lower capacity of the column.

It is advisable to inject as small a volume as possible (and reproducible) in a solvent of low eluotropic strength. Otherwise, retention on the stationary phase is incomplete and many compounds will elute partly in the flow-through. Furthermore, separation could be seriously disturbed, which results in unsymmetrical peak shapes and altered retention times. Figure 1 shows the extracted ion chromatograms, which correspond to the molecular ion of 4-glucopyranosyloxybenzoyl choline, a secondary metabolite identified in methanolic seed extracts, injected in either 2 µL 80% methanol (a) or 2 µL 10%

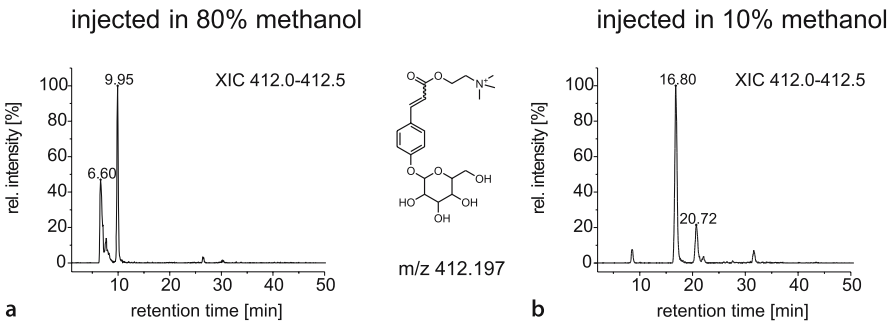


Fig. 1. Influence of solvent on the retention and separation. Extracted ion chromatograms (XIC 412.0–412.5) showing the altered retention behaviour of 4-glucopyranosyloxybenzoyl choline from a seed extract upon injection in different injection solvent mixtures: **a** 80% methanol – a fraction of the metabolite elutes in the flow-through (t_R = 6.60 min); **b** 10% methanol – diastereomers are retained on the column and baseline-separated (t_R = 16.80, t_R = 20.73 min)

methanol (b) following separation on C18 phase with hydrophilic end-capping. In case of an injection in 80% methanol most of the compound is eluted without any retention, whereas upon injection in 10% methanol the compound is retained on the stationary phase and both diastereomeres (probably *cis/trans*-isomers) are baseline separated.

A second major problem for the coupling of LC to MS was initially the incompatibility of commonly used mobile phases with MS. Therefore, non-volatiles such as phosphate ions or the frequently used ion-pairing agent trifluoroacetic acid (TFA), which causes ion suppression due to the extreme ionization capacity of the mother ion, had to be replaced with formate or acetate buffers. The organic component of the mobile phases is most frequently acetonitrile, sometimes methanol. Using classical reversed phase material (RP-18, 3 or 4 μm) as stationary phase, acceptable peak shapes for most of the compounds in leaf and root extracts could be achieved. In general peak widths of about 0.20–0.35 min for a 15-cm column with 3 μm particle size were observed. Application of a C18 phase with hydrophilic end-capping provides better separation for early eluting analytes. In particular, aromatic amino acids and biogenic amines show a considerably improved retention behaviour.

Concerns about the feasibility and reliability of LC-MS-based metabolite profiling have been raised repeatedly (Fiehn 2002; Fernie et al. 2004; Kell 2004). These concerns are mostly referring to the fact that electrospray ionization is prone to matrix effects. This term summarizes two phenomena potentially compromising quantification: (i) reduction or enhancement of ion signals caused by the sample matrix, and (ii) interferences from co-eluting molecules (Matuszewski et al. 2003). Matrix components that are non-volatile can have dramatic effects on the ion signal of an analyte. Mechanistically this effect is not fully understood. It is likely caused by competition between an analyte and non-volatile matrix components for access to the droplet surface in the spray and for reaction with ions formed during the ionization process (Niessen 1999a; Matuszewski et al. 2003; Manini et al. 2004). Thus, reproducibility of a quantification can be compromised, a potential problem that is further aggravated when diverse samples (= matrices) are analyzed. It is important to note, however, that matrix effects have predominantly been observed in cases where there was little chromatographic separation. Run times and column lengths were reduced because the coupling to MS/MS supposedly guaranteed highly selective detection (Matuszewski et al. 2003). It is obvious, however, that good separation prior to ionization is essential to reduce the impact of matrix effects and to minimize ion suppression – especially when highly complex metabolite mixtures are analyzed. The fewer analytes elute from the column simultaneously, the better the chances are of efficient and reproducible electrospray ionization and detection of a particular analyte. Thus, optimal separation is of paramount importance and, consequently, the use of very long monolithic columns for the liquid chromatography has been proposed (Tolstikov et al. 2003). In capillary LC, particle size and column diameter severely restrict the length of the column because of the backpressure build-up. At the same time,

a certain minimum flow rate has to be maintained in order to obtain a stable electrospray. In our experience, therefore, it is not feasible to use columns much longer than 20 cm unless larger particles are used. We found, however, that the 3 µm–15 cm design is sufficient to achieve very good separation. Figure 1b shows as a selected example for the CapLC resolution the base-line separation of the two diastereoisomers of 4-glucopyranosyloxybenzoyl choline (below we will present and discuss an assessment of matrix effects in our metabolite profiling scheme).

A great advantage of ESI is its ability to provide soft ionization. Nevertheless, fragmentation can easily be induced in one of the higher-pressure regions of the ion passageway from the source into the mass analyzer (Fig. 2a). Three potentials which determine the opposite processes of declustering and focusing vs collision induced dissociation (Fig. 2b) can be varied on a QSTAR Pulsar and have to be optimized for the profiling in terms of mass signal yield and appropriate distribution. First we analyzed two model compounds, namely hirsutin (Fig. 2c) and rutin (Fig. 2d), which are prone to give in-source fragments and tried to optimize the intensities of the quasi-molecular ions by systematically ramping both declustering potentials and the focusing potential. For both hirsutin and rutin the quasi-molecular ions $[M+H]^+$ reached their maximum between 40 and 50 V for DP1 and 10 and 15 V for DP2. The effects of the focusing potential on the maxima of the breakdown curves were of minor importance. However, optimization was necessary (FP = 220 V, data not shown). It should be clearly stated that, in principle, for every analyte, such an optimization has to be done to get the full sensitivity, but since a highly complex mixture of mostly unknown compounds is analyzed, compromises have to be made. To get the optimal value of DP1 for a profiling experiment, we measured the same methanolic leaf extract (n = 4) with different DP1 values between 15 and 60 V and analyzed the mass signals with regard to its mass-to-charge ratio and signal-to-noise ratio distribution (Fig. 3, left panel and right panel, respectively). We found that in such a simplified analysis the density functions between 30 and 60 V show only minor differences and, thus, a value of about DP1 = 45 V appears to yield the best results concerning high signal-to-noise ratios as well as high mass-to-charge ratios.

In our metabolite profiling platform (von Roepenack-Lahaye et al. 2004) ions are injected into a QTOF system, a hybrid mass spectrometer. Practically, the third quadrupole in a triple quad instrument is replaced in these instruments with a time-of-flight mass analyzer (Chernushevich et al. 2001). Other mass analysis options for coupling to LC are discussed in detail in another chapter (Sumner et al.; this book, Chap. I.2). Likewise, the QTOF system is dealt with by Bino et al. (this book, Chap. I.3). Thus, we will only briefly summarize our experience with QTOF-MS. We routinely use an acquisition period ("scan time") of 2 s. For the deconvolution of data, which is even more vital for LC due to the lower resolution as compared to GC, there are currently two options available to us. The software MetaboliteID (Applied Biosystems) allows one to extract the mass spectra and to generate an output that lists mass peaks with

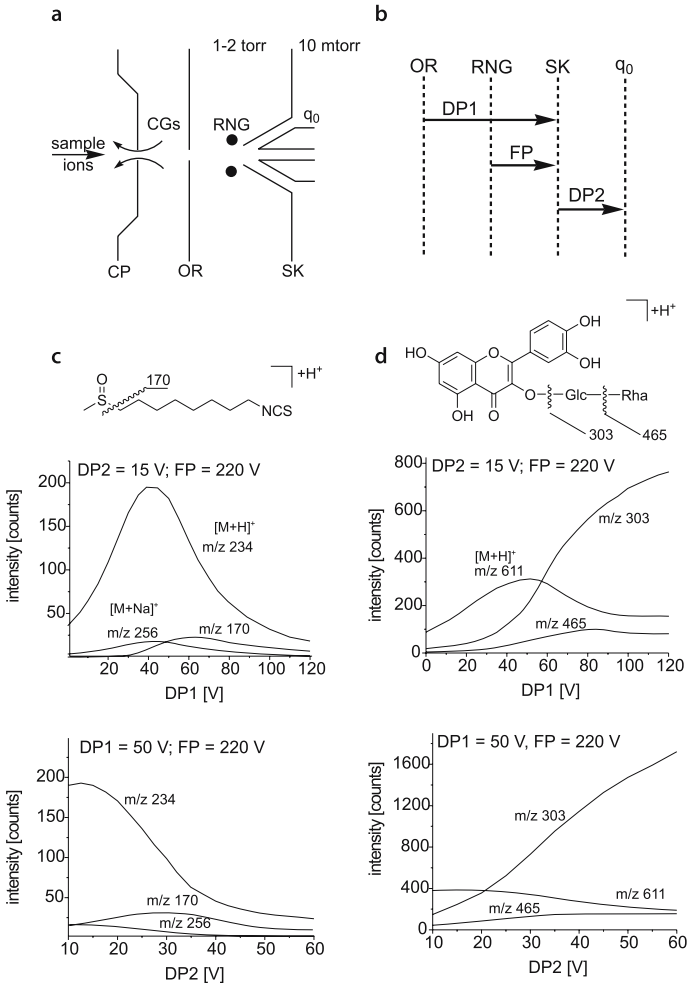


Fig. 2. Effects of ion source potentials on sensitivity and degree of in-source fragmentation: **a** schematic overview of the differentially pumped (evacuated) interface between ion source and mass spectrometer of an API QSTAR Pulsar Hybrid LC/MS system: curtain plate (CP), curtain gas (CGs), orifice (OR), ring (RNG), skimmer (SK); **b** definition of electrical potentials applied in the interfacial region: declustering potential (DP), focusing potential (FP); **c,d** breakdown curves for hirsutin (**c**) and rutin (**d**) obtained in DP1 and DP2 ramping experiments

retention times, accurate mass and intensity. Self-made macros are then needed to normalize, to align peak list and to compare intensities (von Roepenack-Lahaye et al. 2004). These latter steps are covered by the software MetAlign, developed by Arjen Lommen (www.metalign.nl; Tolstikov et al. 2003), which aligns and compares sets of chromatograms to identify differentially abundant mass signals. Reproducibility of the retention times of capillary LC is, in our experience, high enough to allow accurate alignment of chromatograms.

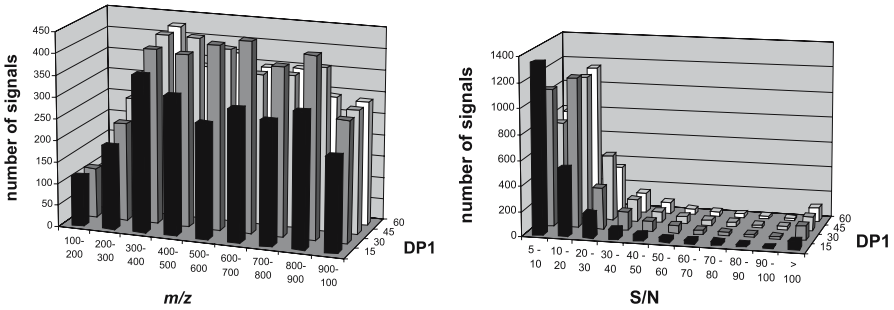


Fig. 3. Effect of declustering potential modulation (DP1 variable, DP2 = 15 V, FP = 220 V) on mass-to-charge ratio (*left panel*) and signal-to-noise ratio (*right panel*) distribution. A methanolic leaf extract was analyzed by CapLC-ESI(+)-QTOF-MS with varying declustering potential 1 settings between 15 and 60 V. Frequencies for different mass-to-charge ratio (m/z) and signal-to-noise ratio (S/N) classes were plotted against DP1 values (n = 4)

3 Potential and Limitations

3.1 Scope of the Analysis

Given the idealistic goal of metabolomics to achieve comprehensive coverage of the metabolome (Oliver et al. 1998), the number of detectable metabolites is an important feature of a metabolite profiling platform. CapLC-ESI-QTOF-MS has great potential because of its sensitivity (Chernushevich et al. 2001). In a single CapLC-MS run analyzed with the deconvolution software MetaboliteID we routinely detect about 1000–2000 mass signals, depending on the extracted material. Running the data through the MetAlign software and applying a signal-to-noise ratio cutoff of 5 gives comparable figures. Similar, albeit somewhat lower numbers (around 700) have been obtained for methanolic *Arabidopsis thaliana* leaf extracts in pilot experiments with monolithic silica columns, coupling of the LC to ion trap MS and deconvolution through MetAlign (Tolstikov et al. 2003).

A mass signal or *m*/*z*, however, does not necessarily represent a metabolite in all cases. Though isotope peaks should be eliminated through the deconvolution, many of the signals are likely to be fragments or adducts so that any given metabolite can in theory give rise to several mass signals. Thus, the number of detectable metabolites is certainly smaller than the number of mass signals and cannot reliably be estimated at this point in time. It is safe to state, however, that several hundred metabolites are routinely detectable in a methanolic extract using the combination of capillary LC and QTOF mass spectrometry. Formation of sodium or potassium adducts, for instance, is not too frequent because of the use of formic acid in the mobile phase.

How large a fraction of the *Arabidopsis* metabolome is covered using this technique? The range of secondary metabolite compound classes detectable

by CapLC-ESI-QTOF-MS in its current state can be assessed by searching in the profiles for members of the various groups of metabolites known to occur in *Arabidopsis thaliana*. A recent compilation listed six biosynthetic classes (d'Auria and Gershenzon 2005): nitrogen-containing compounds, phenylpropanoids, benzenoids, polyketides such as flavonoids, terpenes and fatty acid derivatives. Metabolites of five of these classes can clearly be detected

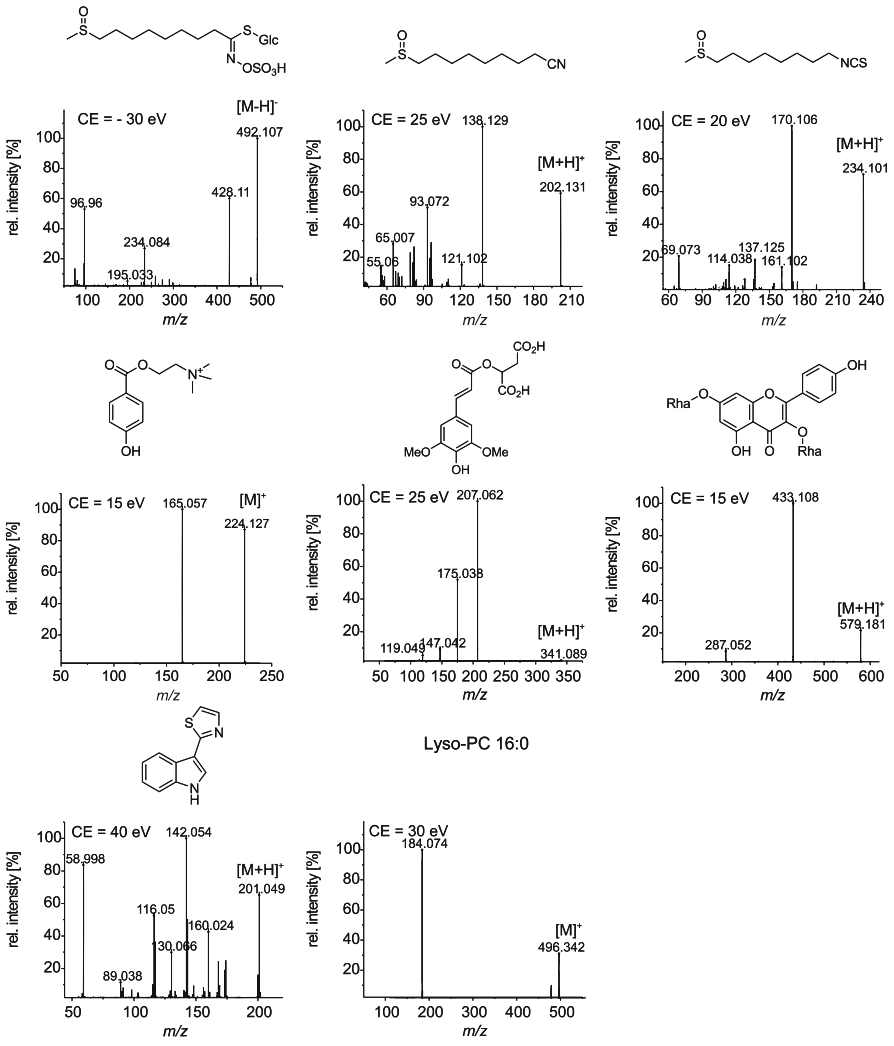


Fig. 4. Most of the known biosynthetic classes of *A. thaliana* secondary metabolites are detectable by CapLC-ESI-QTOF-MS. CapLC-ESI(+/-)-CID-MS spectra of representative metabolites detected in methanolic extracts of different *A. thaliana* tissues such as leaves, seeds and roots (for details on the biosynthetic classes and representative metabolites see text)

by CapLC-ESI-QTOF-MS. Figure 4 displays CID-MS spectra of representatives that we identified in *Arabidopsis thaliana* extracts. Intact glucosinolates can easily be detected by ESI in negative ion mode in nearly all tissues of *Arabidopsis thaliana*. Typical hydrolysis products of glucosinolates like isothiocyanates and nitriles can be detected by ESI(+). Examples are 8-methylsulfinyloctylisothiocyanate (hirsutin) and 8-methylthiononanitrile as well as several homologs. Furthermore, biosynthetic precursors of glucosinolates, like desulfoglucosinolates and thiohydroxamic acids, have been identified in certain cases. Indole-derived secondary metabolites such as the phytoalexin camalexin represent further nitrogen containing compounds that can be detected. Notably, methanolic root extracts contain a huge variety of indole derivatives. Furthermore, ascorbigens and glutathione-indole conjugates, which result from trapping reactions of the hydrolysis products of indole glucosinolates with several nucleophiles, could also be detected. As representatives of the phenylpropanoids, typical esters such as sinapoyl malate in leaves and sinapoyl choline in seeds could be specified. In particular, methanolic seed extracts show a wealth of different choline esters (unpublished observations). Besides the corresponding substituted cinnamoyl cholines various hydroxylated/methoxylated benzoyl cholines could be detected. Other choline containing compounds like differentially substituted phosphatidyl cholines have been identified in seed extracts, too. From the class of flavonoids the major flavonols kaempferol and quercetin and their glycosides could be detected either in positive or negative ion mode. Saccharide composition and aglycon structure can be determined by means of MS^2 and pseudo-MS^3 (product ion spectra derived from in-source CID fragments) experiments.

In conclusion, of the biosynthetic classes known to occur in *Arabidopsis thaliana*, all but one (terpenes) can be detected. Judging from this comparison, CapLC-ESI-QTOF-MS achieves a very good coverage of secondary metabolism. In addition, many primary metabolites such as amino a cids and oligopeptides can be analyzed. This assessment is based on data obtained in positive-ion mode. There is a potential to improve further the reach by also measuring in the negative-ion mode. We found that it is also possible to acquire reliable data in the negative mode without changing the mobile phase, albeit with a significantly reduced mass signal yield.

3.2 Quantification

Proper quantification of an analyte requires optimization of extraction, sample preparation, chromatography and detection, as well as the availability of a pure standard – which would ideally be isotopically labeled – that can be used to calibrate the signal. A standard also allows one to determine the dynamic range of a metabolite in question and to search for possible matrix effects by performing recovery experiments (Birkemeyer et al. 2005). Obviously it is extremely difficult for an unbiased profiling of mostly unknown compounds

to meet the criteria for accurate and reproducible quantification. Standards are available for only a subset of the detected metabolites and this subset is particularly small for capillary LC-MS which targets hundreds of low abundance and species-specific metabolites. At the same time, it is inconceivable that all the required standards can be synthesized. Thus, any metabolomics approach comes at a cost of reduced precision (Trethewey et al. 1999) and has to be assessed critically and improved continuously with respect to accuracy of quantification.

For electrospray ionization there is inherently no correlation between signal strength and abundance when comparing different analytes because ionization efficiency is molecule-dependent. Decisive for quantitative profiling is, however, whether such a correlation exists for any given analyte and what the boundary minimum and maximum detector signals are, i. e. how wide the dynamic range is. In the absence of appropriate standards, serial dilution experiments are a way to assess linearity at least for the mass signals that are sufficiently strong. Our initial results demonstrated for a subset analyzed in detail that there is indeed a good correlation over the almost two orders of magnitude that were tested (von Roepenack-Lahaye et al. 2004). Still, because of the analyte dependency, there is a need to gather information continuously for more and more metabolites and to use these data for the quantification. As a first approximation an extensive set of reference compounds for different compound classes – if at all available – should be used. Eventually, the fascinating concept of mass isotopomer ratio analysis through in vivo labeling with ^{13}C (Birkemeyer et al. 2005) might at some point in the future allow internal standardization of profiling even with multicellular organisms.

The above-mentioned potential of LC-MS for severe matrix effects makes rigorous validation a necessity (Fernie et al. 2004). Matrix effects are again dependent on the analyte and the extract or the nature and origin of the biological sample. Ways to assess these indirectly for an analyte in question are spiking experiments with different matrices (Matuszewski et al. 2003). In order to obtain an estimate of matrix effects at the profiling scale, we performed, for instance, a series of mixing experiments. From previous data we know that leaf and root extracts are fundamentally different in their composition (von Roepenack-Lahaye et al. 2004). Leaf/root extracts were either diluted with 80% methanol or with equal amounts of root/leaf extracts and analyzed ($n = 4$). We focused attention on ions with a strong signal so that they would also be reliably detectable in dilutions. Also, we made sure that signal intensities were in the dynamic range; 45 m/z values eluting between 15 and 45 min were selected. Figure 5 shows the ratios of "signal after dilution with methanol" to "signal after dilution with root/leaf extract". A ratio above 1 is indicative of signal enhancement through the other extract, a ratio below 1 shows signal suppression. One can see in Fig. 5 that ion suppression occurs more frequently than enhancement. Of the 90 m/z measured in different dilutions, 76 showed ratios of 1 ± 0.4 (equals about two times the technical variation) and are therefore considered to be reliably quantifiable. In most experimental set-ups

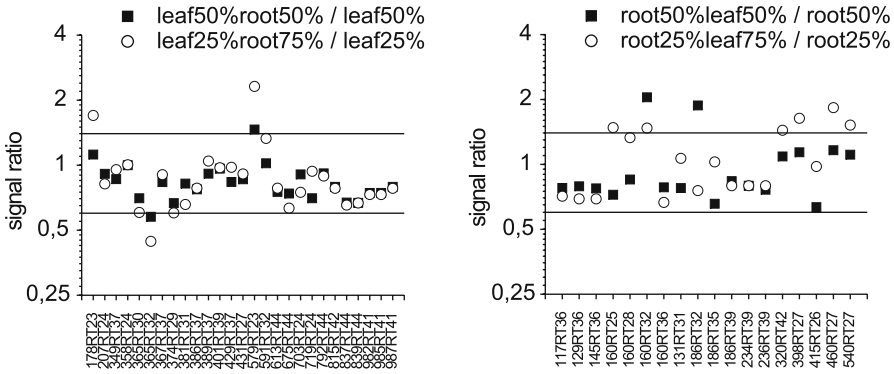


Fig. 5. Evaluation of matrix effects through mixing of extracts of different origin. Signal ratios for mass signals in crosswise matrix-diluted and solvent-diluted leaf and root extracts obtained by CapLC-ESI(+)-TOF-MS measurements. A total of 45 mass signals were analyzed in two different mixtures each. A ratio of "signal after dilution with methanol" to "signal after dilution with root/leaf extract" above 1 is indicative of signal enhancement through the other extract, a ratio below 1 of signal suppression. Most of the mass signals showed a ratio of 1 ± 0.4 (equals about two times the technical variation). This threshold of ±40% is indicated by *vertical lines*

matrix effects will probably be smaller than in this pilot experiment because matrices will be less diverse than a root and leaf extract. Information obtained by analyses such as these can be used to weigh the data obtained in a profiling experiment and to add "confidence tags" to each metabolite. Factored into such "confidence tags" should also be the results on the dynamic range and the degree of variability observed over many experiments. We conclude that the sensitivity and range of CapLC-ESI-QTOF-MS clearly make it a powerful approach for the identification of qualitative differences between samples but that – as predicted by Chernushevich et al. (2001) – quantitative analysis is also feasible provided analyte-dependent effects can be detected and corrected for. All available data suggest that overall the analytical variation is smaller than the biological variation – as is the case for the more established metabolomics approaches (Dunn et al. 2005).

3.3 Identification of Unknown Metabolites

As stated in the introduction, a major potential advantage of LC-MS-based profiling is the multitude of options to obtain structural information on unknown compounds. This is of paramount importance given, for instance, the conservatively estimated 5000 metabolites in *Arabidopsis thaliana* of which maybe 500 are annotated today (Bino et al. 2004). The first piece of information on unknowns is the accurate mass that can be obtained by TOF-MS with a deviation of only 5–10 ppm even in complex matrices (von Roepenack-Lahaye et al. 2004; Ibanez et al. 2005). Based on this, potential elemental compositions

can be calculated. As recently proposed by Ibanez et al. (2005), the usually large number of possibilities can be reduced by calculating theoretical isotopic percentages for all possible elemental compositions and by comparing these to the experimental data. A further significant reduction of formulae can then be achieved through the second and third layer of structural information, in-source fragments and CID-MS spectra (see Fig. 4). The very high mass accuracy of QTOF instruments also applies to product ion scans (Chernushevich et al. 2001). With the information on accurate masses of precursor and product ions, databases can be searched. Obviously, the success rate of this approach is determined not only by the performance of the analysis but also by the availability of databases. In this respect the current situation is far from being satisfactory and future joint efforts will hopefully result in a significant improvement (Bino et al. 2004). One should add, however, that in the end identification will in most cases be tentative and further validation will be required (Bino et al. 2004). Also, discrimination between isomers is not possible without standards.

4 Conclusions and Outlook

Our experience with respect to the potential of this technique for metabolomics can in part be validated by taking a look at recent applications of LC-MS in general and of CapLC-ESI-QTOF-MS in particular. In occupational toxicology the superiority of LC-MS-MS with respect to sensitivity is now being exploited for the determination of trace and ultra-trace amounts of biomarkers of exposure (Manini et al. 2004). Quantification of low-abundance molecules in highly variable complex matrices is considered feasible, provided that precautions such as those outlined above are taken. Also, many novel metabolites have been identified and minor metabolic routes for well-known occupational hazards have been uncovered (Manini et al. 2004). Similarly, the mass accuracy and sensitivity of QTOF-MS coupled to liquid chromatography is now being applied to the elucidation of unknown environmental micro-contaminants in, for instance, water samples (Ibanez et al. 2005). Studies of this kind face challenges similar to those of metabolomics experiments. The emerging picture is that CapLC-ESI-QTOF-MS can be routinely applied (von Roepenack-Lahaye et al. 2004; Bino et al. 2005) and has high potential not only for the identification of selected molecules but for a highly sensitive, robust metabolite profiling that achieves very good coverage of the metabolome. Obviously this technology will be undergoing continuous validation and improvement. Developing the profiling is an iterative process. Any progress made with respect to the availability of standards or reference compounds, the identification of metabolites, the linear range or the possible matrix effects for a particular mass signal has to be used to increase further the accuracy of quantification. Also, protocols for extraction, selective enrichment of metabolites and chromato-

graphic separation have to be tailored for specific questions so that a toolbox of different profiling schemes becomes available to fully exploit the power of CapLC-ESI-QTOF-MS.

Despite the current limitations – comparatively high cost and the lack of LC-MS spectral databases – this profiling approach most likely will contribute substantially to cataloguing the metabolome of *Arabidopsis thaliana* and other systems that are under investigation as models or economically important species. Also, it will help to elucidate biological functions of metabolites and will greatly facilitate the identification of enzyme substrates and products through the systematic analysis of mutants and the correlation with transcript and protein data (Hirai et al. 2005). Extensive data matrices will allow one to unravel metabolic and regulatory networks, especially in secondary metabolism.

References

Abian J, Oosterkamp AJ, Gelpi E (1999) Comparison of conventional, narrow-bore and capillary liquid chromatography mass spectrometry for electrospray ionization mass spectrometry: Practical considerations. J Mass Spectrometry 34:244–254

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci 9:418–425

Bino RJ, Ric de Vos CH, Lieberman M, Hall RD, Bovy A, Jonker HH, Tikunov Y, Lommen A, Moco S, Levin I (2005) The light hyperresponsive high pigment-2dg mutation of tomato: alterations in the fruit metabolome. New Phytol 166:427–438

Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J (2005) Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling. Trends Biotechnol 23:28–33

Chernushevich IV, Loboda AV, Thomson BA (2001) An introduction to quadrupole-time-of-flight mass spectrometry. J Mass Spectrom 36:849–865

D'Auria JC, Gershenzon J (2005) The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. Curr Opin Plant Biol 8:308–316

Dixon RA (2001) Natural products and plant disease resistance. Nature 411:843–847

Dunn WB, Bailey NJ, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130:606–625

Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5:763–769

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48:155–171

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey R, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nature Biotechnol 18:1157–1161

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22:245–252

Halket JM, Waterman D, Przyborowska AM, Patel RK, Fraser PD, Bramley PM (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. J Exp Bot 56:219–243

Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in

Arabidopsis by integration of metabolomics and transcriptomics. J Biol Chem (epub ahead of print)

Ibanez M, Sancho JV, Pozo OJ, Niessen W, Hernandez F (2005) Use of quadrupole time-of-flight mass spectrometry in the elucidation of unknown compounds present in environmental water. Rapid Commun Mass Spectrom 19:169–178

Kell DB (2004) Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol 7:296–307

Manini P, Andreoli R, Niessen WM (2004) Liquid chromatography-mass spectrometry in occupational toxicology: a novel approach to the study of biotransformation of industrial chemicals. J Chromatogr A 1058:21–37

Matuszewski BK, Constanzer ML, Chavez-Eng CM (2003) Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. Anal Chem 75:3019–3030

Niessen WM (1999a) State-of-the-art in liquid chromatography-mass spectrometry. J Chromatogr A 856:179–197

Niessen WM (1999b) Liquid chromatography-mass spectrometry, 2nd edn. Dekker, New York

Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16:373–378

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23:131–142

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Taylor LP, Grotewold E (2005) Flavonoids as developmental regulators. Curr Opin Plant Biol 8:317–323

Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. Anal Chem 75:6737–6740

Trethewey RN, Krotzky AJ, Willmitzer L (1999) Metabolic profiling: a Rosetta Stone for genomics? Curr Opin Plant Biol 2:83–85

Von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S (2004) Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol 134:548–559

Wilm MS, Mann M (1994) Electrospray and taylor-cone theory, Dole's beam of macro-molecules at last? Int J Mass Spectrom Ion Processes 136:167–180

Winkel-Shirley B (2001) Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiol 126:485–493

I.6 NMR Spectroscopy in Plant Metabolomics

J.L. Ward and M.H. Beale[1]

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is one of the most powerful and widely used structural analysis techniques available to the analytical phytochemist and continues to be the technique of choice for unknown structure determination. As a technique for plant metabolomics it benefits from the fact that it is non-compound class selective and non-sample destructive. NMR spectra contain a wealth of accurate qualitative and quantitative information regarding the components of a sample. Whilst measurements of the ^{1}Hs are the most commonly used for metabolomic studies, analysis of the ^{13}Cs has also been employed. However, the low sensitivity of ^{13}C-NMR (due to its lower natural abundance and magnetogyric ratio) prevents its routine use for large numbers of complex extracts. General disadvantages of the sensitivity of NMR (relative to mass spectroscopy) and overlapping signals can be largely overcome, for single compounds and partially fractionated mixtures, by use of instruments with higher field strength magnets (600 MHz or greater) or by the use of modern cryoprobes. In very complex mixtures, such as crude plant extracts, that contain compounds at widely differing concentrations, sensitivity and overlapping signals are problematic for traditional 1D-NMR spectral interpretation. Techniques to deal with overlapping signals, such as 2D-J-resolved spectroscopy, can provide reconstructed 1D spectra that are simplified by the absence of proton–proton coupling (Viant 2003). However, in this review we concentrate on high-throughput 1D ^{1}H-NMR and demonstrate how the technique in combination with chemometrics has become well-established as a plant metabolomic screen. We also discuss how application of 2D- and hyphenated NMR techniques can provide further solutions to the sensitivity and deconvolution problems associated with analysis of complex natural product mixtures. For more general reviews in the use of NMR in plant sciences the reader is directed to Roberts (2000), Ratcliffe et al. (2001) and Ratcliffe and Shachar-Hill (2001, 2005).

[1] The National Centre for Plant and Microbial Metabolomics, Rothamsted Research, West Common, Harpenden, Herts. AL5 2JQ, UK, e-mail: Jane.ward@bbsrc.ac.uk, mike.beale@bbsrc.ac.uk

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

2 High-throughput Screening by 1D ^{1}H-NMR

A key advantage in the use of ^{1}H-NMR spectroscopy in metabolomic screens is the robustness of the technique such that any compound that is soluble in the solvent of choice will be detected, providing that it contains hydrogen atoms. Furthermore, integration of signals from different compounds is absolutely quantitative and is truly representative of the relative concentrations of those compounds. The fact that NMR reliably detects most compounds present gives the technique a clear advantage in screening and fingerprinting applications over mass spectroscopic techniques, which are beset by problems caused by variable ionisation of different types of compounds.

Methodologies for metabolomic screening of plant extracts by NMR spectroscopy are based on the large body of work carried out in the biomedical area, particularly on plasma and urine in relation to disease biomarkers and drug metabolism. Much of this work was done by the prolific research group of Nicholson, Lindon and Holmes at Imperial College, London (Nicholson et al. 1999; Lindon et al. 2000, 2001, 2004; Bollard et al. 2005).

In plant metabolomics, solvent extraction of metabolites from tissue is necessary. Apart from the experimental design aspects where decisions on the plant numbers, growth, and tissue type have to be made, key choices influencing the range of metabolites detected must be made. These are (i) whether to use fresh or freeze-dried tissue and (ii) the polarity of the solvent to be used. On the whole, better quality NMR spectra are obtained by directly extracting freeze-dried tissue with deuterated NMR solvents. Most published work utilises polar solvents (aqueous buffers, perchloric acid, or methanol-water mixtures), although chloroform has also been utilised (Choi et al 2004a). Figure 1a depicts the ^{1}H-NMR spectrum of a deuterated water-methanol extract of *Arabidopsis thaliana.* The spectrum is typical for most green tissue polar extracts (Charlton et al. 2003; Ward et al. 2003; Choi et al. 2004b,c), and is dominated by signals arising from carbohydrates, amino-acids and organic acids. Polar extracts of other plant tissues such as potato tubers (Defernez et al. 2004), wheat flour (Lewis et al. 2003) and tomato fruits (Le Gall et al. 2003) have similar NMR spectra but contain other features reflecting the concentration of certain metabolites associated with these storage tissues. Exudates from plant roots also contain distinctive metabolites (Fan et al. 1997) and represent an area that has been neglected in the recent upsurge in metabolomic studies.

Classic interpretation of complex NMR spectra such as Fig. 1a is difficult although some 30 or 40 metabolites can be definitively identified by virtue of signals that appear in non-overlapped regions of the spectrum and quantified by integration against the internal reference standard (usually trimethylsilyl-*d*4-propionate). Use of libraries of standard spectra run in the same solvent on the same instrument (or at least on an instrument of the same field strength) is an aid to this interpretation, as is 'spiking' of extracts with pure compounds. In the example shown in Fig. 1b, clear differences in the anomeric proton region in the spectrum of wild-type *Arabidopsis thaliana* and that of two different

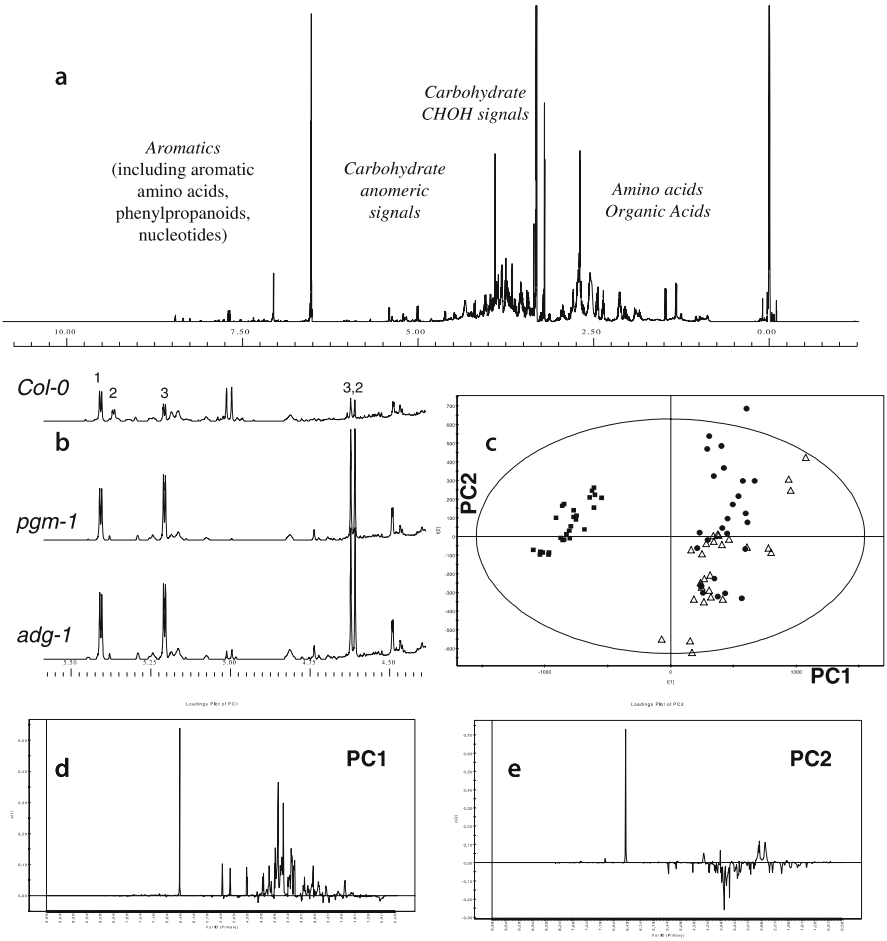


Fig. 1. a 600 MHz ^{1}H NMR spectrum of 4:1 (D_2O:CD_3OD) extract of freeze-dried *Arabidopsis thaliana* Col-0 tissue. **b** Expanded portion of the spectrum featuring the carbohydrate anomeric proton region highlighting differences in carbohydrate concentrations between Col-0 and the starch biosynthesis mutants *pgm-1* and *adg-1*. Labelled peaks: *1*-sucrose, *2*-maltose, *3*-glucose. **c** PCA scores plot illustrating differences observed between Col-0 and the *adg-1/pgm-1* mutants (*filled squares* – Col-0, *open triangles* – *pgm-1*, *filled circles* – *adg-1*). **d** Loadings plot of PC1 depicting the 'spectra' of compounds responsible for differences between Col-0 and the mutants. **e** Loadings plot of PC2 depicting the differences between *adg-1* and *pgm-1*

mutants in the starch biosynthesis pathway (*pgm-1* and *adg-1*) can be seen. These differences can be interpreted in relation to the known function of the enzymes missing in the mutants. However, in most plant metabolomic applications many hundreds of similar spectra are collected. Simultaneous classical interpretation of these numbers of individual spectra is not possible, but the use of chemometrics (see next section) allows the spectroscopist to

focus on compounds that are responsible for differences between individual plants, or populations of plants, and target more detailed analysis to particular compounds or biochemical pathways.

3 Data Analysis

NMR-based metabolomic datasets are very large, both in terms of the number of datapoints per sample (typically 32 k or 64 k), and also the number of samples and resulting spectra acquired (from dozens to thousands, often including replicates). In order to draw conclusions and make comparisons between large numbers of spectra, automated strategies must be employed for the analysis and interpretation of such data once they have been acquired. The literature on data analysis is extensive and will only be briefly discussed here. Interested readers are directed to the review on pattern recognition methods (Lindon et al. 2001) for further coverage of some of the issues.

Data manipulation typically starts with some form of 'bucketing' or 'binning' whereby the spectrum is split into discrete regions (typically between 0.01 and 0.04 ppm in width), which are then integrated to return a list of integral values for each spectrum. Whilst this reduces the resolution of the data, it has the advantage of removing small chemical shift changes due to slight pH variation between samples. Increasingly, work is being carried out using all of the datapoints in the spectrum by employing an algorithm to align the peaks, eliminating any unwanted variation (Stoyanova et al. 2004). NMR data is usually analysed initially using multivariate statistical methods such as Principal Component Analysis (PCA). PCA is a data visualisation method, useful for observing groupings within large datasets. There are a number of commercially available software products that carry out PCA and other related multivariate analyses. One that has been widely used for NMR data is SIMCA-P (Umetrics, Sweden). A PCA model can be displayed in a graphical fashion as a "scores" plot as shown in Fig. 1c. This example compares ^{1}H-spectra collected from polar extracts of wild-type *Arabidopsis thaliana* Col-0 with those from the two mutants in starch biosynthesis (*pgm-1* and *adg-1*). This plot is useful for observing any groupings in the data set and in addition will highlight outliers that may be due to errors in sample preparation or instrumentation parameters etc. Coefficients by which the original variables must be multiplied to obtain the score are called "loadings". Thus, "loading plots" [e. g. Fig. 1d,e] can be used to detect and display the spectral areas responsible for the separation in the data, and can be interpreted as positive and negative NMR spectra of the compounds responsible for the differences between the clusters. The numerical value of the loading of a given variable on a PC indicates how much the variable has in common with that component (Massart et al. 1988). In Fig. 1d, PC1 represents the NMR spectra of compounds differing between wild-types and both mutants, whilst PC2, Fig. 1e, represents the (smaller) difference between the mutants.

When carrying out PCA it is necessary to apply scaling methods to the bucketed data matrices. In NMR spectroscopic data, although the integral values across a spectrum are proportional to concentration and the number of resonances present, the largest resonances would, without scaling, have a dominant effect in multivariate analysis. Before PCA the data can be scaled in different ways. In the covariance matrix method the data are just mean-centred. In the correlation matrix method the data are mean-centred and then the columns (variables) of the data matrix are scaled to unit variance. Covariance matrices are most widely used for NMR data because they have the advantage that the loadings plots retain the scale of the original data and can be compared back to libraries of spectra for assignment. Variable stability scaling (VAST) has recently been described and offers advantages over previously employed scaling methods in terms of the downstream multivariate modelling (Keun et al. 2003). This method weights each variable according to a metric of its stability and can unearth subtle differences between lines against backgrounds of biological variation. However, in the plant research arena, where growth of many identical clones under controlled environment is easily achieved, biological variation can be minimised, for example by pooling many individuals, and thus presents less of a problem.

An alternative method of highlighting differences between sample sets against backgrounds of biological or experimental variation is Orthogonal Signal Correction (OSC) (Gavaghan et al. 2002). This data filtering method can be applied to the scaled data matrix before multivariate analysis, and can, if used carefully, yield insights that are not evident from PCA. Such deeper mining of large data sets for differences against a background of noise can also be obtained by supervised modelling methods such as Partial Least Squares-Discriminant Analysis (Lindon et al. 2001). The use of neural networks to classify spectra has been applied to the study of the herbicide mode of action on maize seedlings (Aranìbar et al. 2001).

4 Two-dimensional NMR

Two-dimensional (2D) NMR experiments, which make use of interactions between NMR-detectable nuclei within a molecule, can be used to increase the spectral resolution and highlight which peaks belong to the same molecule. These experiments generally have much longer acquisition times, posing problems to those researchers wanting to carry out high throughput data collection. Nevertheless, these experiments can be run in automation and generate data useful to the metabolomics researcher and are particularly useful in the assignment of identities to unknown peaks.

The 2D experiments can be split into homonuclear and heteronuclear experiments. Homonuclear experiments examine the correlations between nuclei of the same type (commonly ^{1}H). The TOCSY experiment (Total Correlation

Spectroscopy) describes all interactions in a spin system and therefore is one of the most informative experiments available. The experiment has been used to examine complex matrices such as root exudates and cell extracts (Fan et al. 1997). Heteronuclear correlation experiments – as the name suggests – examine the correlation between two different types of nuclei within a molecule. Indirect experiments such as HMQC (heteronuclear multiple quantum coherence) and HSQC (heteronuclear single quantum coherence) spectroscopy are particularly useful in structure assignment as is HMBC (heteronuclear multiple bond coherence) which can examine long range correlations. A relatively new technique DOSY (diffusion ordered spectroscopy), that is not dependent on analysis of spin–spin coupling, has been applied to analysis of complex mixtures such as liquid foods (Gil et al. 2004). Here resonances are separated in the second dimension by virtue of their diffusion coefficient. This coefficient is governed by molecular size and thus DOSY represents a novel tool to deconvolute overlapping signals and may be particularly useful in plant spectra to assign carbohydrate signals to mono-, di- or higher saccharides.

5 Stable Isotope Labelling

NMR spectroscopy is not restricted to the analysis of ^{1}H signals. ^{2}H, ^{13}C and ^{15}N isotopes, however, have a very low natural abundance, making the detection of these signals difficult. Stable isotope labelling leads to the selective enhancement of some of these signals, providing a powerful method for the scrutiny of metabolic pathways in many organisms including plants (Roberts 2000; Roscher et al. 2000). Stable isotope labelling in plants has been extensively reviewed in recent years (Ratcliffe et al. 2001; Ratcliffe and Shachar-Hill 2001, 2005). ^{13}C is the most useful isotope and there are many suitable precursors including acetate, amino acids, carbohydrates and carbon dioxide. ^{15}N labelling can be achieved using labelled nitrate or ammonium ions whilst D_2O can be used to supply plant tissue with ^{2}H. In the majority of cases, ^{13}C labelling studies involve the use of singularly labelled precursors although multiply labelled precursors can also be used. One of the first applications of this technique in plant metabolism concerned the measurement of ^{13}C–^{15}N bonds using a solid state cross polarisation technique (Schaefer et al. 1981). An important example that demonstrates the use of isotope labelling in plant metabolonics has been published by Kikuchi et al. (2004). Using wild type and ethanol-insensitive mutants of *Arabidopsis thaliana,* labelled with ^{13}C, they were able, using subtractive 2D-HSQC, to isolate and assign only those metabolites that were affected by ethanol treatment.

Isotopic labelling techniques are often used for measuring the fluxes of metabolites through metabolic pathways. The information can be used to establish the identity of the biochemical pathways involved. This type of study has been carried out to investigate metabolic pathways in plants using ^{2}H,

^{13}C and ^{15}N-labelling (Fox et al. 1995; Prabhu et al. 1996; Schleucher et al. 1998). The analysis of stable isotope labelling in pulse-chase and time-course experiments can also provide quantitative information on metabolic fluxes although this is restricted to simple linear pathways that are reasonably close to the entry point of the label into metabolism. For complicated pathways it may be useful to examine the distribution of the label once the system has reached a steady state.

6 Hyphenated NMR

Liquid chromatography-NMR-mass spectrometry (LC-NMR-MS) is arguably the most powerful of the hyphenated techniques available to the phytochemical researcher (Hostettmann and Wolfender 2001; Wolfender et al. 2001). Metabolite profiling using such hyphenated techniques can help to provide clean spectral information on components of a mixture of unknown metabolites in an extract or fraction, leading to a partial or a complete structure determination in a single online experiment. One of the disadvantages of LC-NMR is its lack of sensitivity, which hampers the on-flow measurement of minor metabolites. Another problem is the need to suppress solvent signals from the mobile phase which if left unsuppressed would dominate the spectrum. Signals residing near these solvent peaks may be suppressed together with the solvent signal. This can be a major drawback when dealing with unknown constituents. In these cases the use of sequential analysis using different solvent systems for the LC is necessary.

LC-NMR hyphenation has been a reality for over 20 years (Buddrus and Herzog 1980). However it is only since the improvements in solvent suppression, NMR sensitivity and the use of shielded magnets that the technique has received more widespread recognition. The technique has been successful when applied to plant extracts rich in natural products of relatively low molecular mass. Recent studies have emphasised the value of LC-NMR as a technique for obtaining detailed chemical profiles of species for taxonomic work. For example, using LC-NMR in both on-flow and stop-flow modes, flavones, xanthones and secoiridoids of several *Gentianaceae* taxa have been identified (Wolfender et al. 1997). Vogler et al. (1998) also used LC-NMR in the on-flow mode to identify nine anti-bacterial sesquiterpene lactones from a partially purified extract of *Vernonia fastigiata* (Asteraceae) without the need for isolation of the individual compounds.

LC-SPE-NMR is a relatively new concept with incorporation of an online SPE (solid phase extraction) cartridge to trap an analyte peak prior to introduction into the NMR flow probe. This can be done in automation without interruption of the column flow. An additional advantage of this system is that after drying the cartridge, analytes can be eluted in fully deuterated solvents, reducing the need for solvent suppression. Furthermore multiple trapping on the same

cartridge concentrates the analytes. There are still relatively few publications using this technology although, an application of LC-SPE-NMR to the detection of compounds from oregano was recently reported (Exarchou et al. 2003). Very recently the technology was used for the rapid identification of antioxidants in complex commercial rosemary extracts (Pukalskas et al. 2005). In this work, all major compounds present in the extract were collected on SPE cartridges after their separation and analysed by both NMR and ESI-MS. LC-SPE-NMR using post column solid-phase extraction was also applied to the direct analysis of phenolic compounds in the polar fraction of olive oil (Christophoridou et al. 2005). As well as the identification of simple phenolic acids, lignans and flavonoids the technique enabled the identification of several new phenolic compounds not previously reported as constituents of olive oil.

7 Discussion: Applying NMR to Plant Metabolomics

NMR has proven to be an exceptionally useful tool in animal metabolomics. In plant metabolomics 1D-NMR coupled with multivariate pattern matching serves as an excellent screen to cluster plant lines/treatments by global analysis of the total extractable metabolome. This type of analysis serves as a first pass screen that also gives quantitative data on important abundant metabolites. The clustering information and preliminary metabolite data can then be used to guide more detailed analysis by other techniques. These more targeted techniques include the GC-, LC- and CE-mass spectroscopic techniques discussed elsewhere in this volume, but also include the 2D-NMR, hyphenated-NMR and isotope-labelling techniques described above.

Examples of 1D-NMR-PCA application can be found in several areas, for example, in functional genomics (Cornah et al. 2004; Le Gall et al. 2005), in analysis of ecotypic and cultivar variation (Ward et al. 2003; Frederich et al. 2004), in safety evaluation of GM crops (Noteborn et al. 2000; Charlton et al. 2003; Le Gall et al. 2003; Lewis et al. 2003; Defernez et al. 2004; Manetti et al. 2004), in analysis of the affects of infection (Choi et al. 2004b,c), in classifying mode of action of chemicals (Aranibar et al. 2001) and of course in quality control of food and herbal products (Vogels et al. 1996; Charlton et al. 2002). The scale of this type of screening will increase and new challenges facing researchers in this area concern the construction of databases of fingerprints from mutants and the interfacing of these with similar databases of spectra of pure standards, such that automated interpretation of complex spectra can be performed. Generic plant metabolomic problems such as temporal batch to batch variation caused by machine or chromatographic drift are not generally seen for NMR as instrument drift is minimal. Furthermore, for *Arabidopsis thaliana* at least, biological batch to batch variation in NMR spectra has been eliminated by careful control of growth and experimental procedures (Lewis et al. 2003). Thus the functional genomic goal of a database of electronically

comparable profiles of large collections of gene knockout mutants and other genetic resources may now be achievable.

The potential of 2D- and hyphenated NMR to increase the number of metabolites that can be observed and quantified is yet to be realised. Although throughput will be decreased, technology platforms where selected samples from first pass 1D-NMR-MS-PCA screens are selected for further fractionation by SPE-NMR-MS and LC-SPE-NMR-MS are being put in place. Success with these will inevitably broaden the metabolome coverage, especially when used in parallel with GC-, CE- and LC-MS^n methods.

References

Aranìbar N, Singh BJ, Stockton GW, Ott KH (2001) Automated mode-of-action detection by metabolic profiling. Biochem Biophysl Res Commun 286:150–155

Bollard ME, Stanley EG, Lindon JC, Nicholson JK, Holmes E (2005) NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. NMR Biomed 18:143–162

Buddrus J, Herzog H (1980) Coupling of HPLC and NMR.1. Analysis of flowing liquid-chromatographic fractions by proton magnetic-resonance. Org Magn Reson 13:153–155

Charlton AJ, Farrington WHH, Brereton P (2002) Application of ^{1}H-NMR and multivariate statistics for screening complex mixtures: quality control and authenticity of instant coffee. J Agric Food Chem 50:3098–3103

Charlton A, Allnutt T, Holmes S, Chisholm J, Bean S, Ellis N, Mullineaux P, Oehlschlager S (2003) NMR profiling of transgenic peas. Plant Biotech J 2:27–35

Choi HY, Kim HK, Hazekamp A, Erkelens C, Lefeber AWM, Verpoorte R (2004a) Metabolomic Differentation of *Cannabis sativa* Cultivars using ^{1}NMR spectroscopy and principal component analysis. J Nat Prod 67:953–957

Choi H, Choi HY, Verberne M, Lefeber AMW, Erkelens C, Verpoorte R (2004b) Metabolic fingerprinting of wild type and transgenic tobacco plants by ^{1}H NMR and multivariate analysis technique. Phytochemistry 65:857–864

Choi HY, Tapias Casas E, Kim KH, Lefeber AMW, Erkelens C, Verhoeven JTH, Brzin J, Zel J, Verpoorte R (2004c) Metabolic discrimination of *Catharanthus roseus* leaves infected by *Phytoplasma* using ^{1}H-NMR spectroscopy and multivariate data analysis. Plant Physiol 135:398–2410

Christophoridou S, Dais P, Tseng L-H, Spraul M (2005) Separation and identification of phenolic compounds in olive oil by coupling high-performance liquid chromatorgraphy with post-column solid-phase extraction to nuclear magnetic resonance spectroscopy (LC-SPE-NMR). J Agric Food Chem 53:4667–4679

Cornah JE, Germain V, Ward JL, Beale MH, Smith SM (2004) Lipid utilisation, gluconeogenesis and seedling growth in Arabidopsis mutants lacking the glyoxylate cycle enzyme malate synthase. J Biol Chem 279:42916–42923

Defernez M, Gunning YM, Parr AJ, Shepherd LVT, Davies HV, Colquhoun IJ (2004) NMR and HPLC-UV profiling of potatoes with genetic modifications to metabolic pathways. J Agric Food Chem 52:6075–6085

Exarchou V, Godejohann M, van Beek TA, Gerothanassis IP, Vervoort J (2003) LC-UV-solid-phase extraction-NMR-MS combined with a cryogenic flow probe and its application to the identification of compounds present in Greek oregano. Anal Chem 75:6288–6294

Fan TMW, Lane AN, Pedler J, Crowley D, Higashi RM (1997) Comprehensive analysis of organic ligands in whole root exudates using nuclear magnetic resonance and gas chromatography-mass spectrometry. Anal Biochem 251:57–68

Fox GG, Ratcliffe RG, Robinson SA, Stewart GR (1995) Evidence for deamination by glutamate-dehydrogenase in higher-plants – commentary. Can J Bot 73:1112–1115

Frederich M, Choiu YH, Angenot L, Harnischfeger G, Lefeber AWM, Verpoorte R (2004) Metabolomic analysis of *Strychnos nux-vomica, Strychnos icaja* and *Strychnos ignatii* extracts by ^{1}H nuclear magnetic resonance spectrometry and multivariate analysis techniques. Phytochemistry 65:1993–2001

Gavaghan CL, Wilson ID, Nicholson JK (2002) Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. FEBS Lett 530:191–196

Gil AM, Duarte I, Cabrita E, Goodfellow BJ, Spraul M, Kerssebaum R (2004) Exploratory applications of diffusion ordered spectroscopy to liquid foods: an aid towards spectral assignment. Anal Chim Acta 506:215–223

Hostettmann K, Wolfender JL (2001) Applications of liquid chromatography/UV/MS and liquid chromatography/NMR for the online identification of plant metabolites. In: Tringali C (ed) Bioactive compounds from natural products-isolation, characterisation and biological properties. Taylor and Francis, London, pp 31–68

Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Holmes E, Lindon JC, Nicholson JK (2003) Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. Anal Chim Acta 490:265–276

Kikuchi J, Shinozaki K, Hirayama T (2004) Stable isotope labelling of *Arabidopsis thaliana* for an NMR-based metabolomics approach. Plant Cell Physiol 45:1099–1104

Le Gall G, Colquhoun IJ, Davis AL, Collins GJ, Verhoeyen ME (2003) Metabolite profiling of tomato (*Lycopersicon esculentum*) using ^{1}H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. J Agric Food Chem 51:2447–2456

Le Gall G, Metzdorff SB, Pedersen J, Bennett RN, Colquhoun IJ (2005) Metabolite profiling of *Arabidopsis thaliana* (L.) plants transformed with an antisense chalcone synthase gene. Metabolomics 1:181–198

Lewis J, Baker JM, Beale MH, Ward JL (2003) Metabolite profiling of GM plants: the importance of robust experimental design and execution. In: Nap JP, Atanassov A, Stiekema WJ (eds) Genomics for biosafety in plant biotechnology. NATO science series I, 359. IOS Press, Amsterdam, pp 47–57

Lindon JC, Nicholson JK, Holmes E, Everett JR (2000) Metabonomics: metabolic processes studied by NMR spectroscopy. Concepts Magn Reson 12:289–320

Lindon JC, Holmes E, Nicholson JK (2001) Pattern recognition methods and applications in biomedical magnetic resonance. Prog Nucl Magn Reson Spectrosc 39:1–40

Lindon JC, Holmes E, Nicholson JK (2004) Toxological applications of magnetic resonance. Prog Nucl Magn Reson Spectrosc 45:109–143

Manetti C, Bianchetti C, Bizzari M, Casciani L, Castro C, d'Ascenzo G, Delfini M, di Cocco ME, Lagana A, Miccheli A, Motto M, Conti F (2004) NMR-based metabonomic study of transgenic maize. Phytochemistry 65:3187–3198

Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kauffman L (1988) Chemometrics: a textbook. Elsevier, Amsterdam

Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29:1181–1189

Noteborn HPJM, Lommen A, van der Jagt RC, Weseman JM (2000) Chemical fingerprinting for the evaluation of unintended secondaty metabolic changes in transgenic food crops. J Biotech 77:103–114

Prabhu V, Chatson KB, Abrams GD, King J (1996) C-13 nuclear magnetic resonance detection of interactions of serine hydroxymethyltransferase with C1-tetrahydrofolate synthase and glycine decarboxylase complex activities in Arabidopsis. Plant Physiol 112:207–216

Pukalskas A, van Beek TA, de Waard P (2005) Development of a triple hyphenated HPLC-radical scavenging detection-DAD-SPE-NMR system for the rapid identification of antioxidants in complex plant extracts. J Chromatography A 1074:81–88

Ratcliffe RG, Shachar-Hill Y (2001) Probing plant metabolism with NMR. Annu Rev Physiol Plant Mol Biol 52:499–526

Ratcliffe RG, Shachar-Hill Y (2005) Revealing metabolic phenotypes in plants: inputs from NMR analysis. Biol Rev 80:27–43

Ratcliffe RG, Roscher A, Shachar-Hill Y (2001) Plant NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 39:267–300

Roberts JKM (2000) NMR adventures in the metabolic labyrinth within plants. Trends Plant Sci 5:30–34

Roscher NJ, Kruger NJ, Ratcliffe RG (2000) Strategies for metabolic flux analysis in plants using isotope labelling. J Biotechnol 77:81–102

Schaefer J, Skokut TA, Stejskal EO, McKay RA, Varner JE (1981) Estimation of protein-turnover in soybean leaves using magic angle double cross-polarization N-15 nuclear magnetic-resonance. J Biol Chem 256:1574–1579

Schleucher J, Vanderveer PJ, Sharkey TD (1998) Export of carbon from chloroplasts at night. Plant Physiol 118:1439–1445

Stoyanova R, Nicholls AW, Nicholson JK, Lindon JC, Brown TR (2004) Automatic alignment of individual peaks in large high-resolution of spectral data sets. J Magn Reson 170:329–335

Viant MR (2003) Improved mthods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Commun 310:943–948

Vogels JTWE, Terwel L, Tas AC, van den Berg F, Dukel F, van der Greef J (1996) Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques. J Agric Food Chem 44:175–180

Vogler B, Klaiber I, Roos G, Walter CU, Hiller W, Sandor P, Kraus W (1998) Combination of LC-MS and LC-NMR as a tool for the structure determination of natural products. J Nat Prod 61:175–178

Ward JL, Harris C, Lewis J, Beale MH (2003) Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. Phytochemistry 62:949–957

Wolfender JL, Rodriguez S, Hostettmann K, Hiller W (1997) Liquid chromatography/ultra-violet/mass spectrometric and liquid chromatography/nuclear magnetic resonance spectroscopic analysis of crude extracts of *Gentianaceae* species. Phytochem Anal 8:97–104

Wolfender JL, Ndjoko K, Hostettmann K (2001) The potential of LC-NMR in phytochemical analysis. Phytochem Anal 12:2–22

I.7 Hetero-nuclear NMR-based Metabolomics

J. Kikuchi[1,2,3] and T. Hirayama[2,3,4,5]

1 Introduction

Novel methods for measurement of living systems are making new breakthroughs in life science. In the era of the metabolome (analysis of all measurable metabolites), a mass spectrometry (MS)-based approach is considered to be the major technology (Aharoni et al. 2002; Fiehn 2002; Sumner et al. 2003), whereas a nuclear magnetic resonance (NMR)-based method is frequently regarded as a minor technology due to its low sensitivity. However, we intend to strengthen the NMR-based approach, using advantages of NMR measurement, such as high quantification, non-invasive measurements, localized in vivo spectroscopy, selectivity of nuclear environments, and validity of structure analysis of diverse biomolecules including stereo-isomers. Attractive NMR-based metabolic analyses can be achieved by uniform stable isotope labeling of organisms allowing the application of multi-dimensional NMR experiments that have been used in protein structure determination (Kikuchi et al. 2004; Kikuchi and Hirayama 2005). Using these novel methods, the dynamic molecular networks inside cells and tissues will be dissected.

2 Historical Aspects of NMR Studies of Plant Metabolism

The history of NMR has been sharpened by a succession of major technological and methodological advances, including greatly enhanced sensitivity due to improvements in electronic devices, probe design, high-field superconducting magnets, the field/frequency stability to allow multi-scan averaging, and also the development of pulse Fourier-transform methods, significant progress in data handling facilities, and the development of multi-dimensional NMR

[1] RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045 Japan, e-mail: kikuchi@psc.riken.jp
[2] International Graduate School of Arts and Sciences, Yokohama City University, 1-7-29 Suehiro, Tsurumi-ku, Yokohama, 230-0045 Japan
[3] CREST, Japan Science and Technology Agency, 4-1-8 Hon-cho, Kawaguchi, 332-0012 Japan
[4] Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, 230-0045 Japan
[5] Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, 305-0074 Japan

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

(Ernst 1992; Claridge 1999). NMR spectroscopy provides many new insights into the physiology of higher plants. The evolution of this particular application of NMR can be traced back to the ground-breaking ^{13}CNMR studies using magic-angle spinning methods (Schaefer and Stejskal 1976). The subsequent developments of the technique and its applications have been charted at regular intervals in the review literature, and although not as widely exploited as its proponents might wish, NMR is now becoming an established technique in the armory of plant biochemists.

3 ^{1}H-NMR-based Metabolomics

NMR signals are highly reproducible, and quantitative assessment of each metabolite in a sample is therefore guaranteed. In contrast, MS-signals are sometimes less quantitative due to problems of matrix effects ("ion suppression" or "ion enhancement") (Mei et al. 2003; Mallet et al. 2004). Because NMR is a nondestructive technique, it is easy to combine NMR analysis with a complementary technique such as gas chromatography/MS or liquid chromatography/MS (Corcoran and Spraul 2003; Ott et al. 2003). In contrast to these applications in which numerous specific metabolites can be identified in complex mixtures, other investigators have addressed the question of whether computer-aided comparisons of the ^{1}H NMR spectra of partially fractionated extracts can yield statistically meaningful metabolic fingerprints of the extracted tissue. Using this approach, it was possible to show that there were minimal compositional differences between certain transgenic and non-transgenic tobacco varieties, but only after accounting for the substantial effects of external factors (Choi et al. 2004).

4 Use of Stable Isotope Labeling Technique to Enable Monitoring of the Dynamic Movement of Metabolites

NMR signals can be detected from the nuclei of many isotopes; ^{1}H, ^{13}C, ^{15}N and ^{31}P are the most widely used for biological NMR spectroscopy (Ratcliffe et al. 2001). For carbon the relevant magnetic isotope is ^{13}C. Its natural abundance is only 1.11%, contributing to the considerably lower sensitivity for ^{13}C NMR than for ^{1}H NMR. Accordingly, the application of ^{13}C NMR in unlabeled systems is largely confined to the detection of the most abundant metabolites, such as the organic solutes that accumulate in response to salt stress or certain secondary metabolites (Ratcliffe and Sachar-Hill 2001). Indirect detection techniques such as ^{13}C-hetero-nuclear single quantum coherence (HSQC) pulse sequence increases the sensitivity of the experiments (Vuister and Bax 1992). An example of this approach can be found in an analysis of alkaloid biosynthesis in vivo (Hinse et al. 2003).

The nitrogen atom has two magnetic isotopes, ^{14}N and ^{15}N, and both can be useful for the detection of metabolites in vivo and in extracts. The practicality of detecting the naturally abundant (99.63%) ^{14}N isotope was first demonstrated in root tissues and subsequently in vivo ^{14}N NMR has mainly been used for the analysis of ammonium and nitrate. The extremely low natural abundance of the ^{15}N isotope (0.037%) rules out the detection of unlabeled metabolites, but after labeling with [^{15}N]ammonium or [^{15}N]nitrate it is possible to use in vivo ^{15}N NMR to detect amino acids, as well as certain secondary products. NMR methods are relatively insensitive, so only signals from compounds present at relatively high levels (concentrations of at least 10 μmol/L) can be detected in spectra (Krishnan et al. 2005). Since metabolic engineering often results in the accumulation of relatively high concentrations of metabolites, this insensitivity is often not as restrictive for compound detection and identification as it is in other areas of biochemistry.

5 Approach for Hetero-nuclear NMR-based Metabolomics

In recent years, hetero-nuclear NMR methods and their spectral editing technologies have developed rapidly. For example, careful selection of window functions and base-line corrections of two dimensional (2D)-spectra yielded improved signal dispersion and line shapes of cross peaks permitting clear subtraction 2D-spectra, a technically difficult and time-consuming procedure using conventional 1D-NMR technology (Deferenz and Colquhoun 2003). With the methodology used in recent protein NMR studies, differences in the molecular composition between wild type and mutant strains can be easily quantified. Therefore, we think advanced technologies in NMR analysis combined with stable isotope labeling are useful tool for metabolomic analysis. We report here stable isotope labeling experiments in *Arabidopsis* using carbon or nitrogen, two of the largest components of all organic compounds (Kikuchi et al. 2004; Kikuchi and Hirayama 2005). Figure 1 shows the basic concept of NMR-based plant metabolomics proposed in this study. The NMR-based approach has an advantage when comparing different samples. Spectral subtraction between different mutants or stimuli enables metabolite levels between different samples to be quantified.

5.1 Example of Hetero-nuclear NMR Experiments *In Vitro*

To demonstrate the usefulness of the NMR method (metabolomics), the metabolite profile of an ethanol-hypersensitive mutant of *Arabidopsis, gek1* (Hirayama et al. 2004), was analyzed (Fig. 2a). *Arabidopsis* seedlings were grown for two weeks on agar plates (see above) and treated with 0.5% ethanol or water for 10 h. Figure 2b shows the ^{1}H–^{13}C HSQC spectra of ethanol or water-treated wild type *Arabidopsis* extracts. The subtraction spectra were obtained

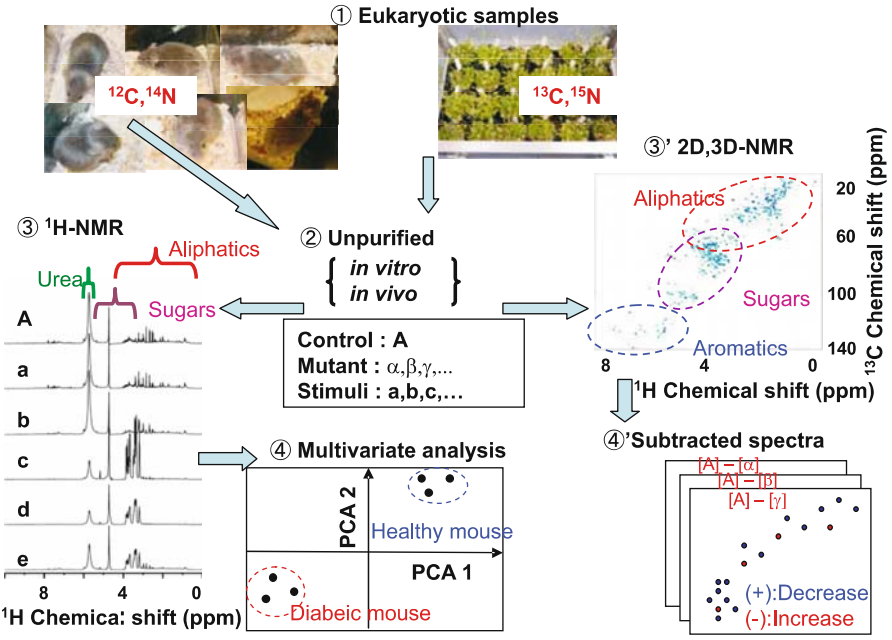


Fig. 1. Comparison of ordinal metabolomics approach (*left*: PCA-based) and our hetero-nuclear NMR metabolomics approach (*right*: multi-dimensional NMR-based)

by subtracting the spectrum of an ethanol-treated sample from that of a water-treated sample. Figure 2c shows the subtraction spectra from the wild-type (WT) (left) or the *gek1* (right) samples. The subtraction of measured spectra generates virtual NMR spectra that highlight compounds that are different between samples. In the present case, the subtraction spectra clearly show that upon ethanol treatment, glutamic acid is synthesized de novo in both the WT and the mutant, consistent with the previous observation that ethanol is converted into amino acids and lipids rapidly in plant tissues (Mellema et al. 2002; Rawyler et al. 2002). In addition, the ethanol-hypersensitive *gek1* mutant synthesized proline and γ-amino butylic acid (GABA) de novo, two compounds that have been reported to accumulate in cells under abiotic stresses such as drought and salinity (for reviews, Hare et al. 1998; Shelp et al. 1999). The assignments of these compounds were possible by comparing both ^{1}H and ^{13}C chemical shifts independently obtained with corresponding commercial reagents. Since ^{13}C-NMR chemical shifts are sensitive to differences in chemical structure but insensitive to the surrounding environment such as solvent effects (Kikuchi and Asakura 1999), the 2D-HSQC type spectra offer exceptionally useful information for assignment of individual chemical groups. From this point of view, construction of a database of 2D-HSQC spectra of main metabolites will enhance the NMR metabolomics.

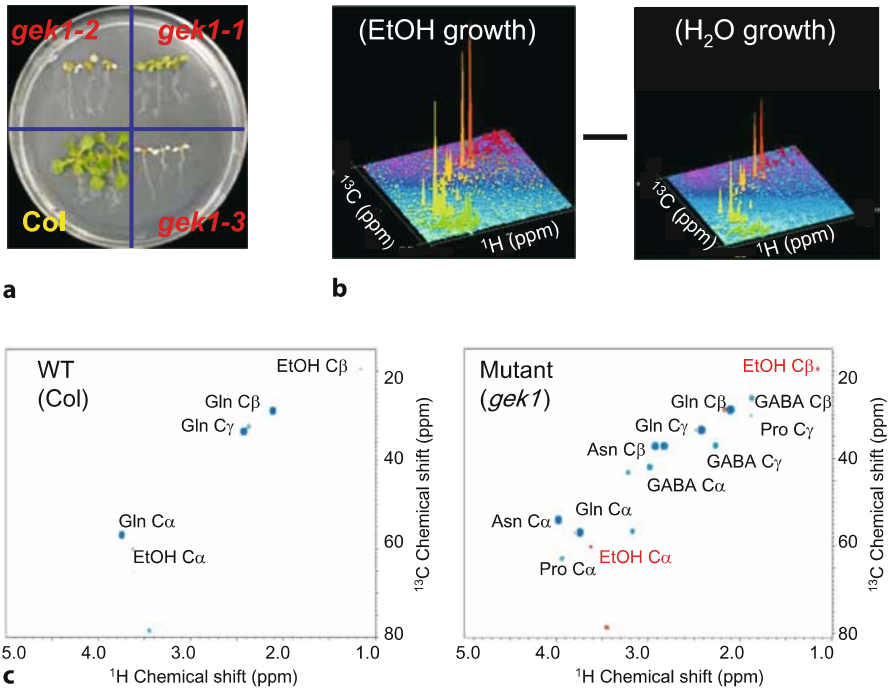


Fig. 2. a Example of how spectral subtraction can be used to differentiate environmental stress responses between WT and *gek1* (Hirayama et al. 2004) mutant. **b** For NMR spectroscopy, 5 mg of freshly frozen samples were heated with 0.5 mL H_2O and centrifuged at 15,000 *g* for 5 min to remove insoluble fractions. After adding 50 µL of ^2H_2O for NMR lock, supernatants were transferred into 5-mm NMR tubes. The spectra were measured on a Bruker DRX-500 spectrometer equipped with a ^1H inverse probe with triple axis gradient. A total of 200 complex f1 (^{13}C) and 1024 complex f2 (^1H) points were recorded with 64 scans per f1 increment. The spectral widths were 12,000 Hz and 8400 Hz for f1 and f2, respectively. **c** To quantify the signal intensities, a Lorentzian-to-Gaussian window with a Lorentzian line width of 10 Hz and a Gaussian line width of 15 Hz was applied in both dimensions, prior to Fourier transformation. A fifth order polynomial baseline correction was subsequently applied in the f1 dimension (Kikuchi et al. 2002). The indirect dimension was zero-filled to 2048 points in the final data matrix. NMR spectra were processed using NMRPipe software (Delaglio et al. 1995). Quantitative 2D-spectral subtraction was accomplished by editing a macro program of the NMRPipe software. Signal assignments are highlighted next to the corresponding cross peaks

5.2 Example of Hetero-nuclear NMR Experiments In Vivo

^{15}N uniformly labeled Arabidopsis seeds can be obtained from plants fed with a nutrient solution containing $^{15}NO_3$ as the sole nitrogen source. Using such seeds, the first ^1H–^{15}N HSQC-type NMR (Bodenhausen and Ruben 1980; Grzesiek and Bax 1993) in vivo experiments in plants were performed (Kikuchi et al. 2004). Figure 3 shows the development of the ^1H–^{15}N HSQC spectrum measured in living ^{15}N-labeled seeds that was induced by soaking the dry

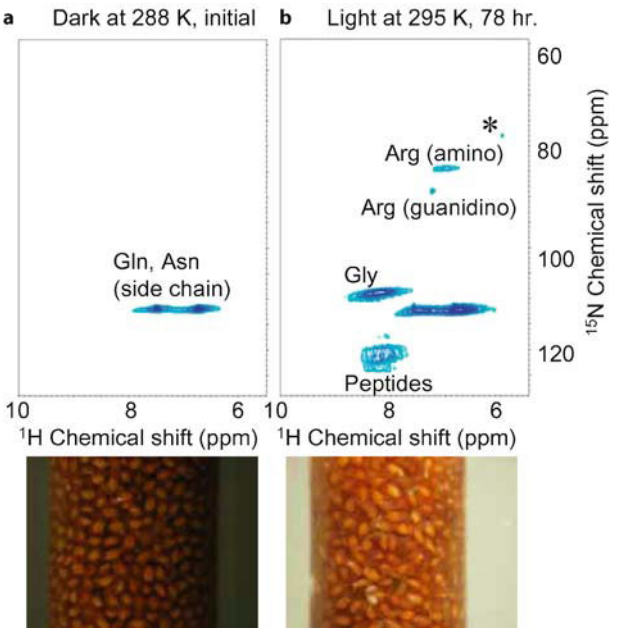


Fig. 3. Development of the ^{1}H–^{15}N HSQC spectrum measured in living ^{15}N-labeled seeds that was induced by soaking the dry seeds in water (*pictures shown at the bottom*). A total of 128 complex f1 (^{15}N) and 1024 complex f2 (^{1}H) points were recorded with 96 scans per f1 increment. The spectral widths were 4500 Hz and 8400 Hz for f1 and f2, respectively. Two spectra: **a** dark at 0 h; **b** light at 78 h are shown for comparison. Signal assignments *are highlighted* next to the corresponding cross peaks

seeds in water. Using in vivo measurement, dynamic movement of metabolites can be observed. In this case, at the initial stage just after water absorption into dried ^{15}N seeds, all cross peaks (especially those corresponding to peptide backbones) were very broad due to slow molecular motion in the dried seeds (Fig. 3a). After 12 h of imbibition, the line shapes of cross peaks, especially those corresponding to the glycine backbone and the side-chains of glutamic acid and asparatic acid, started to sharpen due to the enhanced molecular motion caused by the increasing water content. The temperature shift from 4 to 22 °C at 72 h of imbibition, and light illumination which started at 78 h of imbibition, both of which accelerate germination, further sharpened all cross peaks and enhanced the peptide backbone signal dramatically (Fig. 3b).

6 Prospects for the Future

As described above, NMR techniques possess an advantage over common analytical methods because they simultaneously provide information on the

concentrations of numerous compounds as well as their spatial distribution. Therefore, NMR offers useful methodology for metabolomics. At this moment, however, there are several issues to be resolved before we can utilize the full power of NMR measurement in metabolomics. First, the sensitivity of NMR is rather lower. This disadvantage is being overcome with the progress in NMR technology. The sensitivity of a spectrometer scales as the 7/4th power of the static magnetic-field, and our group has developed the highest magnetic field (21.2 Tesla) used in biomolecular studies (Kiyoshi et al. 2004). In addition, NMR signal-to-noise (S/N) ratios can be substantially improved by cooling the NMR radio frequency detector and preamplifier . We are currently developing a 4.5-K cryogenetically cooled probe for the 920-MHz NMR spectrometer (Yokota et al. 2004). The increase of S/N gain is expected to be 8-fold, corresponding to a 64-fold reduction of the NMR acquisition time. The ^{1}H-^{13}C HSQC spectrum (shown in Fig. 1) recorded by the 500-MHz spectrometer exhibited 477 cross peaks identified by the NMRPipe software (Delaglio et al. 1995), corresponding to 100–200 metabolites at concentrations over 10^{-6} mol/L. However, theoretically the 920-MHz spectrometer equipped with a 4.5-K cryogenically cooled probe will be able to detect metabolites at concentrations as low as 10^{-8} to 10^{-9} mol/L. Furthermore, S/N gain by the cryogenetically cooled probe is significantly enhanced in low dielectric solvents (Horiuchi et al. 2005). In other words, the ^{1}H–^{13}C HSQC spectra recorded by 64 scans with the 500-MHz spectrometer (shown in Fig. 2) will be taken by only one scan with equivalent S/N ratio for the same sample but with higher resolution due to the higher magnetic field. Second, to facilitate the identification of metabolites in samples, a database of 2D HSQC spectra of known metabolites is required. We have just started to construct such a database. Once developed, metabolite analyses will be conducted with simultaneous quantification and metabolite identification. Furthermore, recent solid-state NMR methods will facilitate the study of insoluble metabolites such as starch, cell-wall components, and biomembranes (Kikuchi et al. 2000). Thus, the use of isotope labeling together with newly developed NMR technologies open a new avenue for plant metabolomics.

Acknowledgements. This work was supported in part by RIKEN GSC Internal Collaborations (No. 830-56625), by CREST (No. A88-54366), Japan Science and Technology Agency to J.K., T.H. We also acknowledge Grants-in-Aid for Scientific Research (No. 15710171, to J.K.; No. 15570045, to T.H.) from the Ministry of Education, Science, Sports and Culture of Japan.

References

Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. OMICS 6:217–234

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189

Choi H-K, Choi YH, Verberne M, Lefeber AWM, Erkelens C, Verpoorte R (2004) Metabolic fingerprinting of wild type and transgenic tobacco plants by ^{1}H NMR and multivariate analysis technique. Phytochemistry 65:857–864

Claridge TDW (1999) High-resolution NMR techniques in organic chemistry. Elsevier Science, London, UK

Corcoran O, Spraul M (2003) LC-NMR-MS in drug discovery. Drug Discov Today 8:624–631

Defernez M, Colquhoun IJ (2003) Factors affecting the robustness of metabolite fingerprinting using ^{1}H NMR spectra. Phytochemistry 62:1009–1017

Delaglio F, Grzesiek S, Vuister G W, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Ernst RR (1992) Nuclear magnetic resonance Fourier transform spectroscopy. Angew Chem Int Ed Engl 31:805–823

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48:155–171

Grzesiek S, Bax A (1993) The importance of not saturating H_2O in protein NMR. Application to sensitivity enhancement and NOE measurements. J Am Chem Soc 115:12593–12594

Hare PD, Cress WA, van Staden J (1998) Dissecting the roles of osmolyte accumulation during stress. Plant Cell Environ 21:535–553

Hinse C, Richter A, Provenzani J, Stöckgt J (2003) In vivo monitoring of alkaloid metabolism in hybrid plant cell cultures by 2D cryo-probe NMR without labeling. Bioorg Med Chem 11:3913–3919

Hirayama T, Fujishige N, Kunii N, Iuchi S, Shinozaki K (2004) A novel ethanol hypersensitive mutant of *Arabidopsis*. Plant Cell Physiol 45:703–711

Horiuchi T, Takahashi M, Kikuchi J, Yokoyama S, Maeda H (2005) Effect of dielectric properties of solvents on the quality factor for a beyond 900 MHz cryogenic probe model. J Magn Reson 174:34–42

Kikuchi J, Asakura T (1999) Use of ^{13}C conformation-dependent chemical shifts to elucidate the local structure of a large protein with homologous domains in solution and solid state. J Biochem Biophys Method 38:203–208

Kikuchi J, Hirayama T (2005) Novel methods for uniform stable isotope labeling in plant and animal systems for a hetero-nuclear NMR based metabomics. 1st Int Metabol Meeting

Kikuchi J, Williamson MP, Shimada K, Asakura T (2000) Structure and dynamics of photosynthetic membrane-bound proteins in *Rhodobacter sphaeroides*, studied with solid-state NMR spectroscopy. Photosyn Res 63:259–267

Kikuchi J, Iwahara J, Kigawa T, Murakami T, Okazaki T, Yokoyama S (2002) Solution structure determination of the two DNA-binding domains in the *Shizosaccharomyces pombe* Abp1 protein by a combination of dipolar coupling and diffusion anisotropy restraints. J Biomol NMR 22:333–347

Kikuchi J, Shinozaki K, Hirayama T (2004) Stable isotope labeling of Arabidopsis thaliana for a hetero-nuclear NMR-based metabolomics approach. Plant Cell Physiol 45:1099–1104

Kiyoshi T, Maeda H, Kikuchi J, Ito Y, Hirota H, Yokoyama S, Ito S, Miki T, Hamada M, Ozaki O, Hayashi S, Kurihara N, Suematsu H, Yoshikawa M, Matsumoto S, Sato A, Wada H (2004) Present status of 920 MHz high-resolution NMR spectrometers. IEEE Trans Appl Supercond 14:1608–1612

Krishnan P, Kruger NJ, Ratcliffe RJ (2005) Metabolic finger printing and profiling in plants by NMR. J Exp Bot 56:255–265

Mallet CR, Lu A, Mazzeo JR (2004) A study of ion suppression effects in electrospray ionization from mobile phase additives and solid-phase extracts. Rapid Commun Mass Spectrom 18:49–58

Mei H, Hsieh Y, Nardo C, Xu X, Wang S, Ng K, Korfmacher WA (2003) Investigation of matrix effects in bioanalytical high-performance liquid chromatography/tandem mass spectrometric assays: application to drug discovery. Rapid Commun Mass Spectrom 17:97–103

Mellema S, Eichenberger W, Rawyler A, Suter M, Tadege M, Kuhlemeier C (2002) The ethanolic fermentation pathway supports respiration and lipid biosynthesis in tobacco pollen. Plant J 30:329–336

Ott K-H, Aranibar N, Singh B, Stockton GW (2003) Metabolomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. Phytochemistry 62:971–985

Ratcliffe RJ, Shachar-Hill Y (2001) Probing plant metabolism with NMR. Annu Rev Plant Physiol Plant Mol Biol 52:499–526

Ratcliffe RJ, Roscher A, Shachar-Hill Y (2001) Plant NMR spectroscopy. Prog NMR Spect 39:267–300

Rawyler A, Arpagaus S, Braendle R (2002) Impact of oxygen stress and energy availability on membrane stability of plant cells. Ann Bot 90:499–507

Schaefer J, Stejskal EO (1976) C-13 Nuclear magnetic resonance of polymers spinning at magic angle. J Am Chem Soc 98:1031–1032

Shelp BJ, Bown AW, McLean MD (1999) Metabolism and functions of gamma-aminobutyric acid. Trends Plant Sci 4:446–452

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Vuister GW, Bax A (1992) Resolution enhancement and spectral editing of uniformly ^{13}C enriched proteins by homonuclear broadband ^{13}C–^{13}C decoupling. J Magn Reson 98:428–435

Yokota H, Okamura T, Ohtani Y, Kuriyama T, Takahashi M, Horiuchi T, Kikuchi J, Yokoyama S, Maeda H (2004) 4.5 K cooling system for a cryogenically cooled probe of a 920 MHz NMR. Adv Cryo Eng 49:1826–1833

II.1 Bioinformatics Approaches to Integrate Metabolomics and Other Systems Biology Data

B. Mehrotra and P. Mendes[1]

1 Introduction

To understand the functioning of cells fully it is important to unravel the roles of genes and their products. The study of gene transcripts (transcriptomics) and proteins (proteomics) is progressing rapidly through the use of microarrays and mass spectrometry. Additionally, cells contain numerous other organic molecules not directly encoded in the DNA, the metabolites, which are critical for cell function. Knowledge about metabolites is crucial for an understanding of most cellular phenomena (Weckwerth 2003; Fernie et al. 2004; Kell 2004). Metabolomics is an emerging field consisting of the study of metabolites at a systems scale. It is similar in objectives to transcriptomics and proteomics; two major goals of metabolomics are the identification of all metabolites in each organism (their metabolomes) and measurements of their dynamics under many different challenges. Integrated approaches combining metabolomics with transcriptomics and proteomics are now underway (e. g. Verhoeckx et al. 2004; Broeckling et al. 2005) and are expected to result in much deeper insights than any of these techniques alone.

Metabolomics shares several characteristics with proteomics and transcriptomics. Like these, it is a technique where large numbers of molecules are profiled simultaneously (though current methods identify only hundreds of metabolites, vs thousands of proteins, and tens of thousands of transcripts). Metabolite profiles are, like transcript and protein profiles, snapshots of the state of a biological sample. Experimental data of all three types are usually dominated by a number of variables (molecules) much larger than samples, posing a hard challenge for data analysis and interpretation. Like proteomics, most metabolomics methods rely on spectroscopies to identify molecules; however, this is done through comparison against standards, rather than by mass fingerprints or sequence. The lack of a concept of "sequence" for metabolites is a major difference between metabolomics and the other two methodologies. Sequence is the key for identification of proteins and nucleic acids in large-scale profiles, but alternative methods must be used for metabolite profiles. De novo identification of metabolites can be done with 2D NMR, but requires considerable amounts of highly purified material, a major obstacle. Tandem mass spectrometry requires smaller amounts of material, but alone is often

[1] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington St., MC 0477, Blacksburg, Virginia 24061, USA, e-mail: mendes@vt.edu

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

insufficient to identify completely unknown metabolites. The alternative to de novo metabolite identification is to construct a library of standard profiles, created with purified metabolites (e. g. Wagner et al. 2003). While large spectral libraries for NMR, IR and mass spectrometry have existed for decades, these are very incomplete compared to the nearly 200,000 known natural products (Buckingham 1994). Additionally, these published spectral libraries are often not accurate enough for clear identifications. In the case of chromatography-mass spectrometry techniques, much better results are obtained if one has constructed a library with one's own equipment, which requires a large investment of time and finance. Thus, metabolite identification is one of the fundamental differences between metabolomics and proteomics/transcriptomics, and it is common to find two thirds or more unidentified metabolites in non-targeted profiles. The second major difference between these techniques is that metabolites have widely different chemical properties, such as polarity, volatility, molecular mass, and chemical reactivity. Comparatively, nucleic acids are very uniform in their properties; proteins, while more diverse than the former, are still approachable by their common properties (such as the amide bonds that can be used to sequence them). Because metabolites vary in their composition and structure, they require many different methods for extraction and separation, and no single existing technique is able to profile all metabolites in a biological sample (Sumner et al. 2003). Comprehensive coverage of the metabolome requires parallel analyses carried out with several different techniques.

Since the turn of the century, systems approaches have regained popularity with biologists. Perhaps this is because the analysis of purified molecules is rapidly approaching its limit, or simply because global experimental analyses have become possible. Either way, it is now recognized that systems biology studies of complex cellular phenomena are sorely needed (Kitano 2002). An increasingly appealing approach consists of experiments that simultaneously monitor the levels of transcripts, proteins, and metabolites, and combine their data to make inferences about the structure and dynamics of the underlying biochemical networks (Mendes 2001; Mendes et al. 2002; Oliver et al. 2002). In order to integrate the diverse and large amount of data generated by such experiments, several statistical and computational methods are required. In particular, it is important that all data be managed in a single database which also keeps track of the intricate details about how the experiments have been designed and the data generated. Ultimately, the data must be traceable backwards to samples and experiments, allowing not only for their interpretation but also for enabling others to replicate the experiments. These data about data are usually known as *metadata* and there have been several attempts at standardizing them. The MIAME standard was proposed for transcriptomics data (Brazma et al. 2001) and received widespread support, including support by prominent journals requiring data to conform to that standard. Similar proposals for proteomics data have been put forward, e. g. PEDRo (Taylor et al. 2003) and MIAPI (Orchard et al. 2004), but these are still under development and have

not yet received the crucial support from the publishing world. Metabolomics is no different, and recently two proposals have been published to define standards for plant metabolomic data and metadata, ArMet (Jenkins et al. 2004) and MIAMET (Bino et al. 2004). Invariably, these attempts build upon the MIAME standard and hopefully will soon allow unequivocal specification of systems biology data. (The existing Systems Biology Markup Language, SBML (Hucka et al. 2003), is a standard for specifying systems biology models, rather than data.)

In the remainder of this chapter we will delineate ongoing efforts in our laboratory pertaining to integrating metabolomics and other functional genomics data. These efforts have arisen from our participation in plant systems biology studies of *Medicago truncatula* and *Vitis vinifera,* and also of the yeast *Saccharomyces cerevisiae*. All of these are large team efforts and we acknowledge all our collaborators for their vital role in these projects (see below).

2 Databases

We are developing a database system, DOME (database for OMEs), which stores functional genomics data originated from microarray measurements of transcripts, 2D-PAGE-QTOFMS protein assays, and GC-MS, LC-MS, CE-MS, and CE-LIF assays of metabolites. Early on it became evident that, in order to analyze all these data sets in an integrated way, they should be stored in a single database. This avoids, by design, the infamous data integration problem of bioinformatics (Davidson et al. 1995), as all data reside in the same schema, and queries can be made across all of them irrespective of their nature. This integration on a single schema was only possible by assuring that all necessary metadata was included and structured in an appropriate way. Figure 1 depicts a high-level overview of the DOME schema.

The main skeleton of the metadata schema consists of a hierarchy of experiment sets, experiments, sampling points, samples, and extracts, following exactly the way in which the biological material is manipulated in the experiments. This metadata schema also allows for easy export of transcript data in a format compatible with MIAME, and hopefully with the future proteomics and metabolomics standards as they stabilize (it is already compatible with PEDRo and ArMet). Linking through metadata is one of the main ways in which one can put together data from metabolomics with microarray and proteomics. For example, one single sampling point (a sample collection time) is attached to a set of perturbed and control replicate transcript levels, protein levels, and metabolite levels, as well as to their respective average values, comparisons and statistical significance. Thus one can relate the metabolite levels at one time in the experiment with the transcript and protein levels and, because the sampling points reflect experiment time, also to the molecular levels of previous or subsequent sampling points.

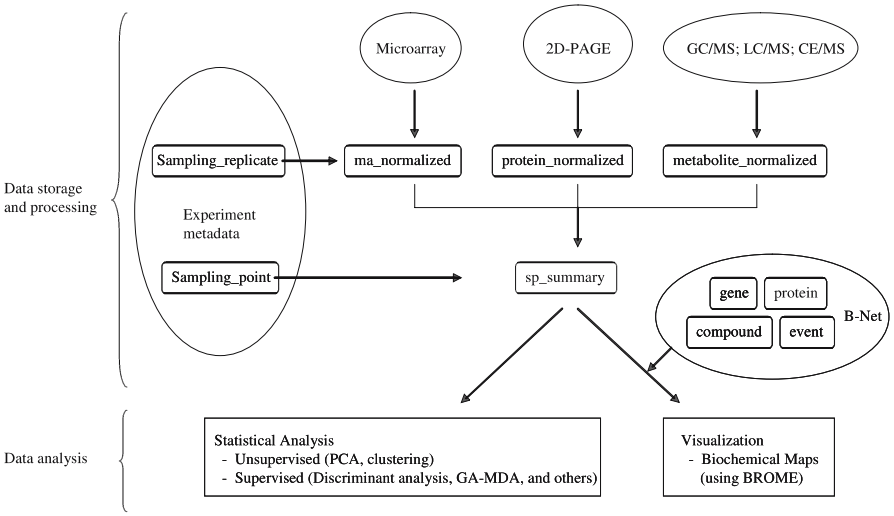


Fig. 1. High level schema of the integrative functional genomics database DOME. Metadata tables are used to provide context to the actual experimental data. Raw data from microarray, 2D-PAGE-MS, and various metabolomics technologies are stored in separate tables. These data are transformed by their appropriate normalization methods, keeping intermediate values, and finally arriving at numerical summaries of each sample (such as means and standard deviations), which are then comparable across all technologies and stored in the sp_summary table. It is the data in sp_summary that is then processed with higher-level statistical analyses or visualizations. It is also at this level that background information about the known molecular biology of the system (B-Net) can be integrated

Another feature that makes the comparison of metabolite with transcript and protein data possible is that each of these data types is first processed in appropriate ways, resulting in data of a similar type. Presently these data are reduced to ratios of levels (usually the level of a molecule in the perturbed state over its level in the control). Before data are available in this state they need to be processed through methods that are different for each technique. For example, the microarray data goes through a series of standardization and normalization procedures that take into account the technical details of microarray technology (Quackenbush 2001). The metabolite GC-MS data are first corrected for sample size, then deconvoluted into a series of peaks, which are identified by the software AMDIS (Davies 1998) relative to an internal standard that was included in the samples. An important issue is that the data should all be represented either in linear or in logarithmic space, not a mixture of the two. It is also important to preserve raw data, because there is always the possibility that in the future better methods to process it will appear; however, because the raw data are not commonly used in the analysis, it is enough to store these offline as long as the database keeps appropriate track of their location.

A third way to relate metabolite data with transcript and protein data is through the use of existing biochemical knowledge. This is, of course, the

traditional way to analyze such comparisons; however, it is usually done by experts in an ad-hoc way. In order for this to be automated into computational procedures, it is necessary to represent this domain knowledge in appropriate schemas. Several biochemical databases exist that partially fulfill our needs (Kanehisa et al. 2004; Krieger et al. 2004), though they fall short in a number of ways (Wittig and de Beuckelaer 2001; Xing Li et al. 2002). In particular, they are rather poor in terms of their coverage of plant secondary metabolism, even AraCyc (Mueller et al. 2003), which is specific for *Arabidopsis*. Thus we created a sub-schema in our database to represent the existing knowledge about the biochemistry of the species in question. Because this can be useful on its own (i. e. independent of the experimental data), it was designed in a manner that allows it to be an autonomous database, and has been named B-Net. B-Net has been populated with gene/transcript information from TIGR's gene indices (Lee et al. 2005) and SGD (Dwight et al. 2004), with protein information from UniProt (Bairoch et al. 2005), metabolite information from LIGAND (Kanehisa et al. 2004) and AraCyc (Mueller et al. 2003), and supplemented with data collected directly from the literature by a team of curators in our laboratory. All of the facts in B-Net are documented for the type of evidence that supports them, using a method generalized from the Gene Ontology's evidence codes (Ashburner et al. 2000). Note that the information imported from external databases is filtered to remove entries that do not represent specific molecular entities. This was particularly important in the case of LIGAND, where groups of molecules are represented alongside individual molecules (e. g. "amino-acids", instead of the specific ones); however only the latter were imported to B-Net. B-Net also classifies entries with gene ontology terms wherever possible.

It is relevant here to highlight two problems that pervade metabolomics databases. The first is the issue of metabolites that are detected but not clearly identified, already mentioned above. These metabolites are sometimes referred to as "unknowns". Despite their identity not being known, a database must distinguish between them, and so these are usually named by the analytical chemists through some ad-hoc scheme. Such names are often attributed in ways that prevent comparison of data between different labs, for example by choosing identifiers that are used in different contexts meaning different things. A negative consequence would be that in two separate experiments two unknown metabolites might receive a common name, even though no one had intended to mean that they were the same molecular entity. In order to overcome this problem, and because our database contains data originated from several laboratories, we have developed a naming scheme that assures that unknown metabolites from different experiments and laboratories are not accidentally named the same thing. This naming convention has been proposed to the community in a recent Opinion article (Bino et al. 2004) and we hope that it becomes adopted by many laboratories, as this is the only way in which it would become useful. Essentially, the name for each unknown metabolite is composed of an identifier for the laboratory, one for the extraction method, another one for the type of analysis carried out, and at least two coordinates

from the analysis. These coordinates are context specific, and could be retention time of a separation, mass-to-charge ratio, chemical shift, wavenumber, etc. Another objective with this naming convention is to allow for future analyses that attempt to establish identity between these unknown metabolites. It is expected that many of these are observed in different studies in separate laboratories, and through the inclusion of the analysis coordinates in their names it becomes easier to recognize that two unknowns may actually be the same molecule. For example, if several studies consistently identified a peak in GC-MS (using the same extraction and analysis parameters) with the same retention time and main ion mass, then it may be that the two are the same metabolite. By assigning names derived from this scheme, it then also becomes possible to create lists of molecular entities that have been observed and not yet identified (a kind of "orphan" list for metabolomics).

Another unresolved issue that is being encountered in our projects is that the same metabolites in a sample might have been observed by more than one technique. The problem that is posed then is which quantification should one chose if they do not agree. This is complicated by the fact that when metabolites appear in an analysis, they may not have been present in the original sample, but instead result from an artifact of the extraction method. Another reason could be that the same metabolite might be present in different locations in the cell, leading to the metabolite being isolated in two separate pools. In the latter case the two pools should both be represented in the database, while in the case of artifacts, one should use only the more accurate quantification. This issue results in a need for careful annotation of metabolomics results, but also requires special structures in the database schema that are capable of representing several pools of a single metabolite.

3 Data Visualization

Scientific data visualization is the activity of displaying properties of a data set that help the human scientist to identify quickly its most important characteristics. This is not a simple problem because it is very hard to identify what would be important for each scientist, and it is as much an issue of the scientific domain as it is of psychology. Nevertheless, there are data properties that are generally sought by a large class of researchers, and visualization software is focused on them. For metabolomics, a frequent way in which biochemists like to visualize data is through the use of maps that depict portions of the biochemical network. Several packages exist that allow for this map-based data visualization (Luyf et al. 2002; Shannon et al. 2003; Thimm et al. 2004; Lange and Ghassemian 2005), and we have also developed one (BROME, BRowser for OMEs) that is coupled to our database system DOME for visualization of metabolomics data with transcriptomics and/or proteomics data. This allows our database to select only a small set of metabolites, enzymes and genes that

are present in a certain map, such that the researcher can then quickly observe their levels organized according to how we believe the biochemistry is organized. This could help in understanding how changes in mRNA or protein levels affect the level of metabolites in a certain pathway or network. However this is not as straightforward as it may seem: the changes in level of mRNA are likely very different from changes in protein levels or changes in metabolite levels. Cells cannot tolerate large changes of many metabolites, while mRNA levels can change widely without much toxicity. Thus, in order to visualize the expression of metabolites, mRNA, and proteins in the same biochemical map, they need to be expressed on different scales, or otherwise normalized to some comparable scale. A problem with thinking about data as part of some biochemical map ("pathway") is that it is likely that molecules in the map are also involved in other interactions not depicted there. Therefore, looking at a particular slice of a network could be highly misleading. It has been shown that the concentrations of metabolites next to each other in a metabolic map do not necessarily have high correlation (Steuer et al. 2003; Camacho et al. 2005), strengthening this point. In order to understand a change in the level of a particular metabolite, it may be more useful to view the expression changes of *all* enzymes (i. e. their protein and mRNA levels) linked with that metabolite. For this we have developed the concept of metabolite neighborhood maps (Xing Li et al. 2002), which are local views of the biochemical network and consist of all the reactions that affect the metabolite of interest, including all the metabolites and enzymes that take part in those reactions. BROME has a large number of maps available, from these neighborhood maps to the nice pathway maps of the KEGG system (Kanehisa et al. 2004).

4 Data Analysis

Analysis of metabolomic data can use the same multivariate statistical methods that are widely used in microarray data analysis. These methods can be either supervised, where each sample or variable (molecule) is associated to an already known class, or unsupervised, where there is no pre-classification of the data (Mendes 2002; Sumner et al. 2003; Goodacre et al. 2004). Unsupervised methods are widely popular, and the most used are principal component analysis (PCA), hierarchical clustering (HCA), k-means clustering, and self-organizing maps (SOM). Unsupervised analyses are mostly guided by the variance and covariance (or correlation) in the data sets, so they are good at finding patterns therein; however nothing guarantees, other than a careful experimental design, that the largest variance is indeed a result of the perturbation rather than other unwanted effects. On the other hand, supervised analyses are guided by the pre-existing knowledge provided by the researcher and so are usually based on discrimination, a property that is more related to the consistency of the members in a class, and the differences between classes. Supervised

methods already demonstrated for metabolomics data are linear discriminant analysis (Raamsdonk et al. 2001; Bundy et al. 2005), discriminant partial least squares (PLS-DA) (Gavaghan et al. 2002; Jonsson et al. 2004), genetic algorithms (Johnson et al. 2003; Goodacre 2005), genetic programming (Allen et al. 2003; Goodacre 2005) and other methods (Goodacre et al. 2000; Shi et al. 2004).

Obviously, much more information could be extracted from systems biology experiments if metabolomics data were analyzed coupled with transcriptomics and proteomics data, but this requires much attention to ensure that the data be comparable, as has been discussed in the previous section (see also Purohit et al. 2004). For example, multidimensional scaling should be preferred to principal component analysis since the former takes into account the different scales of each variable, but the latter does not (at least in its original incarnation that is based on covariance).

We have recently tried to understand how to interpret correlation between metabolites in metabolomics data sets (Camacho et al. 2005) following results from an analysis by Steuer et al. (2003). Briefly, we have found that strong correlation arises when two or more metabolites are in chemical equilibrium, when they conserve a common chemical moiety, when their concentrations are mostly controlled by a single enzyme, or when one of the enzymes that affects their concentration changes with greater magnitude than the other enzymes that also affect them (Camacho et al. 2005). We confirmed that two metabolites next to each other in a biochemical network (e. g. substrate and product of a single enzyme) are not expected to be strongly correlated, in general. These conclusions put the results that have been obtained demonstrating high correlations between metabolite pairs into a systems biology context; often such correlations are only explained at a global level, and through the action of proteins.

Since metabolomics data are best understood when accompanied by other systems biology data, there seems to be a great need for methods that specifically address data integration. As demonstrated by the study of metabolite correlations, methods that take into consideration pre-existing biochemical knowledge are likely to be more effective. Since many metabolomics experiments are carried out through time intervals, methods that consider the time dimension explicitly are likely to be more productive than those that do not.

5 Conclusion

In this chapter we have addressed some of the bioinformatic issues related to metabolomics and its integration within the systems biology framework. We believe that metabolite, transcript, and protein analyses are much more powerful combined than individually. In order to extract maximal benefit from such combined studies, specific bioinformatics support is necessary in the form

of databases, visualization, and data analysis. Ultimately, a full understanding of the underlying phenomena will require an additional layer of computational and theoretical tools, supporting the formulation and evaluation of dynamic models that attempt to represent the biological system. Such models will need to be predictive, but we believe that, much more than that, they need to be explanatory. Within our laboratory we are pursuing several projects in this direction and have a strong interest in combining that approach with the data and informatics systems described here, as have others. This will be a topic of much discussion in the near future and we await it with excitement.

Acknowledgements. We thank our collaborators John Cushman, Grant Cramer, Rick Dixon, Greg May, David Schooley, Vladimir Shulaev, and Lloyd Sumner for the excellent experimental and analytical data collection work in their laboratories. We also thank our colleagues Xing Li and Aejaaz Kamal for their work on the DOME and BROME systems, and Diogo Camacho and Alberto de la Fuente for the metabolite correlation analysis. We acknowledge the generous support of the National Science Foundation's Plant Genome Research Program (awards DBI–0109732 and DBI–0217653).

References

Allen J, Davey HM, Broadhurst D et al. (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nat Biotechnol 21:692–696

Ashburner M, Ball CA, Blake JA et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Bairoch A, Apweiler R, Wu CH et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33:D154–D159

Bino RJ, Hall RD, Fiehn O et al. (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci 9:418–425

Brazma A, Hingamp P, Quackenbush J et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genet 29:365–371

Broeckling CD, Huhman DV, Farag MA et al. (2005) Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. J Exp Bot 56:323–336

Buckingham J (1994) Dictionary of natural products. Chapman and Hall/CRC, London

Bundy JG, Willey TL, Castell RS, Ellar DJ, Brindle KM (2005) Discrimination of pathogenic clinical isolates and laboratory strains of Bacillus cereus by NMR-based metabolomic profiling. FEMS Microbiol Lett 242:127–136

Camacho D, de la Fuente A, Mendes P (2005) The origin of correlations in metabolomics data. Metabolomics 1:53–63

Davidson SB, Overton C, Buneman P (1995) Challenges in integrating biological data sources. J Comput Biol 2:557–572

Davies T (1998) The new Automated Mass Spectrometry Deconvolution and Identification System (AMDIS). Spectroscopy, Europe 10:24–27

Dwight SS, Balakrishnan R, Christie KR et al. (2004) *Saccharomyces* genome database: underlying principles and organisation. Brief Bioinform 5:9–22

Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5:763–769

Gavaghan CL, Wilson ID, Nicholson JK (2002) Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA. FEBS Lett 530:191–196

Goodacre R (2005) Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. J Exp Bot 56:245–254
Goodacre R, Shann B, Gilbert RJ et al. (2000) Detection of the dipicolinic acid biomarker in Bacillus spores using Curie-point pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. Anal Chem 72:119–127
Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22:245–252
Hucka M, Beale M, Fiehn O et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531
Jenkins H, Hardy N, Beckmann M et al. (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol 22:1601–1606
Johnson HE, Broadhurst D, Goodacre R, Smith AR (2003) Metabolic fingerprinting of salt-stressed tomatoes. Phytochem 62:919–928.
Jonsson P, Broadhurst D, Goodacre R et al. (2004) A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem 76:1738–1745
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280
Kell DB (2004) Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol 7:296–307
Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
Krieger CJ, Zhang P, Mueller LA et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32:D438–D442
Lange BM, Ghassemian M (2005) Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry 66:413–451
Lee Y, Tsai J, Sunkara S et al. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33:D71–D74
Luyf AC, de Gast J, van Kampen AH (2002) Visualizing metabolic activity on a genome-wide scale. Bioinformatics 18:813–818
Mendes P (2001) Modeling large scale biological systems from functional genomic data: parameter estimation. In: Kitano H (ed) Foundations of systems biology. MIT Press, Cambridge, MA, pp 163–186
Mendes P (2002) Emerging bioinformatics for the metabolome. Brief Bioinform 3:134–145
Mendes P, de la Fuente A, Hoops S (2002) Bioinformatics and computational biology for plant functional genomics. Rec Adv Phytochem 36:1–13
Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. Plant Physiol 132:453–460
Oliver DJ, Nikolau B, Wurtele ES (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metabolic Eng 4:98–106
Orchard S, Hermjakob H, Julian RK et al. (2004) Common interchange standards for proteomics data: Public availability of tools and schema. Proteomics 4:490–491
Purohit PV, Rocke DM, Viant MR, Woodruff DL (2004) Discrimination models using variance-stabilizing transformation of metabolomic NMR data. Omics 8:118–130
Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2:418–427
Raamsdonk LM, Teusink B, Broadhurst D et al. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotechnol 19:45–50
Shannon P, Markiel A, Ozier O et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
Shi H, Paolucci U, Vigneau-Callahan KE, Milbury PE, Matson WR, Kristal BS (2004) Development of biomarkers based on diet-dependent metabolic serotypes: practical issues in development of expert system-based classification models in metabolomic studies. Omics 8:197–208
Steuer R, Kurths J, Fiehn O, Weckwerth W (2003) Observing and interpreting correlations in metabolomic networks. Bioinformatics 19:1019–1026
Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Taylor CF, Paton NW, Garwood KL et al. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. Nat Biotechnol 21:247–254

Thimm O, Blasing O, Gibon Y et al. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939

Verhoeckx KC, Bijlsma S, Jespersen S et al. (2004) Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis. Int Immunopharmacol 4:1499–1514

Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62:887–900

Weckwerth W (2003) Metabolomics in systems biology. Annu Rev Plant Biol 54:669–689

Wittig U, de Beuckelaer A (2001) Analysis and comparison of metabolic pathway databases. Brief Bioinform 2:126–142

Xing Li X, Brazhnik O, Kamal A et al. (2002) Databases and visualization for metabolomics. In: Harrigan GG, Goodacre R (eds) Metabolic profiling: its role in biomarker discovery and gene function analysis. Kluwer Academic Publ, Boston, pp 293–309

II.2 Chemometrics in Metabolomics – An Introduction

J. Trygg[1], J. Gullberg[2], A.I. Johansson[2], P. Jonsson[1], and T. Moritz[2]

1 Introduction

In the post-genomics era, the use of methodologies that enable transcriptomic, proteomic and metabolomic data to be analysed in detail have revolutionized biological investigations. One of the major advantages with metabolomics investigations compared to traditional target metabolite analysis is that metabolomics data can give an unbiased view of changes in metabolism during environmental, genetic or developmental changes. Instead of tracking only a few metabolites, changes in relative amounts in 300 to 1000 or even more metabolites can be recorded and analysed, covering all major metabolic pathways. This development has accentuated the need to apply and further develop multivariate methodology. Chemometrics (see Eriksson et al. 2001) provides tools to make good use of measured data, enabling practitioners to make sense of measurements and to model quantitatively and produce visual representations of information. Today, chemometrics has grown into a well established data analysis tool in areas such as multivariate calibration, quantitative structure-activity modeling, pattern recognition and multivariate statistical process monitoring and control. Although seemingly diverse disciplines, the common denominators in these applications are that high complexity data tables are generated and that these data tables can be analysed and interpreted by means of chemometric methods.

In chemometrics, there are three basic categories of analysis (Fig. 1):

1. Exploratory analysis (Fig. 1A). This gives an overview of all the data in order to detect trends, patterns or clusters.
2. Classification analysis and discriminant analysis (Fig. 1B), which classifies samples into categories or classes, for example wild-type and mutant.
3. Regression analysis and prediction models (Fig. 1C) are used when a quantitative relationship between two blocks of data is sought. For example, when prediction of growth or fiber properties from mass spectrometry data.

However, in biology, chemometric methodology has still been largely overlooked in favour of traditional statistics. It is not until recently that the

[1] Research Group for Chemometrics; Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, e-mail: johan.trygg@chem.umu.se
[2] Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden, e-mail: thomas.moritz@genfys.slu.se

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

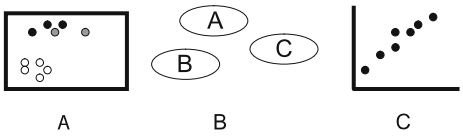


Fig. 1. Overview of the basic categories of chemometrics analysis: **A** overview of data structure; **B** classification and discriminant analysis; **C** regression analysis

overwhelming size and complexity of the 'omics' technologies has driven biology towards the adoption of chemometric methods. Here we will give an introduction to chemometrics and also give examples of why and when chemometrical methodologies should be used.

2 Theory and Methods

2.1 Making Data Contain Information – Design of Experiments

In experimental biology, e. g. when investigating how a number of different environmental factors (e. g. temperature, day length, nutrition) affect different responses such as growth, transcript profiles and metabolite profiles in plants, there is a need to carry out experiments in a systematic way. One way to investigate how the factors affect the plant's responses is to Change One Factor at a Time, i. e. the COST approach. This approach has severe problems: (1) finding optimal conditions for experiments (e. g. method development), (2) unnecessarily many experiments are needed (inefficiency), (3) ignores interaction among variables (lost information) and (4) provides no map over the experimental space.

Design of Experiments (DOE) (Lundstedt et al. 1998) is the methodology of how to conduct and plan experiments in order to extract the maximum amount of information in the fewest number of runs. The basic idea is to devise a small set of experiments, in which all pertinent factors are varied systematically. It is a fundamental tool for planning experiments and making data informative by simultaneously, albeit in a structured way, varying controllable factors (e. g. environmental conditions, instrument settings, experimental procedures) of the studied system. Today they comprise a tool box for virtually any experimental problem.

2.1.1 Stages in the DOE Process

Most of us can only grasp the effect of one factor at a time in our minds, and that often leads us into the inefficient COST approach. We need the mathematics (and the computer) to keep track of the factors and their combinations.

In summary, (1) all factors are varied together over a set of experimental runs, (2) noise is decreased by means of averaging, (3) the functional space is efficiently mapped, interactions and synergisms are seen.

1. What do I want? – formulate question(s) stating the objectives and goals of the investigation. For example identify factors (e. g. temperature, day length, nutrition) and factor ranges (e. g. 15–25 °C, 6–12 h, 1–10 mmol N/L) that affects flowering time.
2. Screening design – finding out a little about many factors. Which factors are the dominating ones in controlling flowering time? Screening designs provide simple models with information about dominating variables, and information about ranges. Pareto's principle states that 20% of the data (factors) account for 80% of the information. Different types of screening designs exist – which one to choose depends on the problem. The most common one is the fractional factorials design (Fig. 2). The full factorial design is a set of experimental runs where every level of a factor is investigated at both levels of all the other factors. It requires $N = 2^k$ number of runs for k factors. Investigating more than five factors with the full factorial design can in some cases become time consuming, i. e. $2^5 = 32, 2^6 = 64, 2^7 = 128$ experiments, etc. Instead, performing a *fractional factorial design* reduces

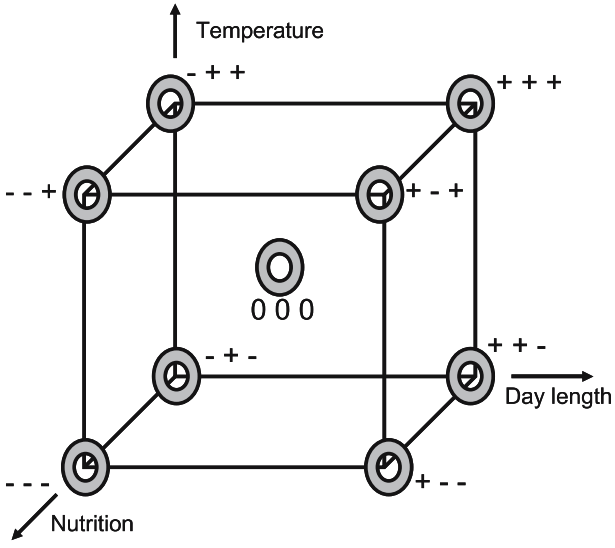


Fig. 2. Example of a full factorial design of experiments (DOE) for investigating how three factors (temperature, day length and nutrition) control flowering time. Varying the three factors at two levels (coded as +/-) requires $2^3 = 8$ experiments + center points. Each experiment according to the design set of experiments is *marked with a circle* in the figure. Evaluating the results from such an experimental design reveals the influence of each of the different factors separately and also any interactions between them. DOE is the only feasible approach to separate cause and effect from each other

that number quickly without the loss of too much information regarding the estimation of factors involved. Fractional factorial design takes advantage of the fact that three-way and higher interactions are seldom significant. It requires only $N = 2^{k-p}$ number of runs for k factors, where p is set manually. For example five factors can be run in only $2^{5-2} = 8$ experiments instead of $2^5 = 32$ experiments compared to the full factorial design. Fractional factorial design takes advantage of the fact that three-way and higher interactions are seldom significant. The downside, of course, for not performing all experiments, is that confounding patterns are present. In other words, the estimated effects are not "pure" but instead mixed with higher degree interaction effects. This loss of information is the prize we need to pay for the reduction of the number of experiments. The degree of confounding is determined by the choice of p.

3. Response surface modeling (RSM) and optimization (few factors) – after screening the factors involved in, e. g. determination of flowering time or derivatization of metabolites, the goal of the investigation is usually to create a valid map of the experimental domain (local space) given by the significant factors and their ranges. This is done with a quadratic polynomial model. The higher order models have an increased complexity, and therefore also require more experiments/factors than screening designs. Different types of RSM designs include Central composite designs, Box Behnken designs and D-optimal designs (see, e. g. Lundstedt et al. 1998 for more information).
4. Robustness testing – in robustness testing of, for instance, an analytical method, the aim is to explore how sensitive the responses are to small changes in the factor settings, e. g. temperature. Ideally, a robustness test should show that the responses are not sensitive to small fluctuations in the factors, that is, the results are the same for all experiments. Robustness testing is usually applied as the last test just before the release of a product or a method. The fractional factorial design is usually applied here.

Plant metabolomic studies typically constitute a set of samples from Arabidopsis wild types and mutants. Assume that these have been subjected to different external conditions such as variation in day length and temperature. Design of Experiments can then be used to select representative samples, related to the biological question we are investigating (how flowering time is affected by temperature, day length, nutrition). An experimental design in three factors can be setup, with factor 1 (temperature), factor 2 (day length), and factor 3 (nutrition). In total, only eight different experiments equal 2^k where $k = 3$ factors are required to explore the experimental space. In addition, a number of replicates, typically three experiments, are added to estimate the noise level. By adding extra experiments, one can investigate more thoroughly the day length and temperature dependence (increase the number of different day lengths and temperatures).

2.2 The Data Table, X-matrix

In plant metabolomics studies, typically a set of samples are characterised using modern instrumentation such as GC/MS, LC/MS or H^1-NMR spectroscopy. The choice of instrument (see Sumner et al. 2003; Dunn et al. 2005) and experimental procedure (Gullberg et al. 2004) are important and largely determined by the biological system and the scientific question. Design of Experiments can here be used to optimize the experimental protocol.

In contrast to a ^{1}H-NMR spectrum, GC/MS and LC/MS data must be processed before multivariate analysis. The reason is the two-dimensional nature (chromatogram/mass spectra) of the data for each sample. For GC/MS data, curve resolution or deconvolution methods are mainly applied for data processing (see, e. g. Halket et al. 1999; Jonsson et al. 2005a). This gives a resolved spectral and chromatographic profile for each detected compound. The 1D multivariate profile used to characterize each sample is made up of the integrated areas of all detected chromatographic peaks. The corresponding mass spectrum and retention index are used for identification purposes (Schauer et al. 2005). For LC/MS data, curve resolution can be applied (e. g. Idborg-Björkman et al. 2003) or a peak detection algorithm that identifies all chromatographic peaks and uses their integrated areas as the multivariate profile characterizing that sample (e. g. Andreev et al. 2003). Another alternative is to sum the chromatographic direction to create a 1D multivariate profile produced by the total intensity over all mass spectral channels (e. g. Allen et al. 2004). Recently, partly alternative methodologies have been applied to GC/MS data (Jonsson et al. 2004, 2005a) and LC/MS data (Jonsson et al. 2005b) where all samples are processed simultaneously and a common set of descriptor variables are extracted.

After, e. g. the GC/MS analysis, we now have a multivariate profile (300–1000 s of variables) for each sample that is a fingerprint of the inherent properties (e. g. phenotype) for each sample. For multiple samples we can therefore construct a two-dimensional data table, an X matrix, by stacking each sample on top of each other. The question is then, how do we go about analysing this multivariate, highly collinear and complex data set? The univariate approach (e. g. student's t-test [Jackson 1991]) is not recommended. It assumes independent variables in X (i. e. more samples than variables) and this creates problems with interpretation, spurious correlations (so called Type I, II errors) and the evident risk of missing information in combinations of variables. Traditional statistical methods (e. g. multiple linear regression, MLR) are also not recommended. They also assume independent variables and have difficulties with noisy data (Eriksson et al. 2001). Instead, multivariate analyses based on projection methods represent a number of efficient and useful methods for the analysis and modeling of these complex data. Projection methods convert the multi-dimensional data table into a low-dimensional model plane, usually consisting of two to five dimensions. Principal component analysis (PCA) (Jackson 1991) and partial least squares (PLS) (Wold et al. 1984) methods are

two widely used methods that can handle incomplete, noisy and collinear data structures.

2.3 Geometrical Interpretation of a Data Table

An easy way to understand and appreciate projection based methods is to translate the data table into a swarm of points in a multi-dimensional space. For a data table or matrix X, with N rows (biological samples) and K columns (e. g. relative amounts of different metabolites), each row (individual sample) can be represented as a point in a K-dimensional space. Its position in this space is given by its coordinates, i. e. its values in each of the K columns. Repeating this for all N rows in a matrix, we have produced a swarm of points in K-dimensional space. Points (samples) that lie close to each other in this multi-dimensional space are more biologically similar to each other than points that lie far apart (dissimilar). Projection methods find a model hyperplanes of much lower dimensionality that closely approximates X, i. e. the swarm of points. Figure 3 gives an overview of how multivariate projection methods work.

2.4 Principal Component Analysis

Principal Component Analysis (PCA) is the workhorse in chemometrics. It is a multivariate projection method designed to extract and display the systematic variation in a data matrix X. The first two *principal components* define a plane, a window into the K-dimensional space. By projecting each of the sample points (in K-dimensional space) onto this two-dimensional sub-space, it is possible to visualize all the samples. The coordinates of each of these samples projected onto this plane are called *scores* T, and they are weighted averages of all X-variables (e. g. metabolites). Hence the visualization of these scores T is called a *score plot*. The score plot is very informative because it gives an overview of all samples in X and how they relate to each other. It may reveal groupings of samples (clusters), trends and outliers (deviating samples). e. g. two genotypes (wild type and mutant) would show up as two distinct clusters of samples, representing wild type and mutant samples respectively. In addition, an experiment that suffered from a broken GC-vial would translate into an unique point in the score plot, i. e. an outlier (Fig. 3).

The score plot allows us to investigate the relation among the samples, but once interesting patterns are found (groupings, outliers etc.), it is possible to understand the reason for this, i. e. what variables (e. g. metabolites) are responsible for this pattern found in the score plot. Hence, there also exists a corresponding plot related to the measured variables (metabolites), i. e. the columns in the X matrix. This plot is known as the *loading plot* P and describes the influence (weight) of the X-variables (metabolites) in the model. An important feature is that directions in the score plot correspond to directions in the

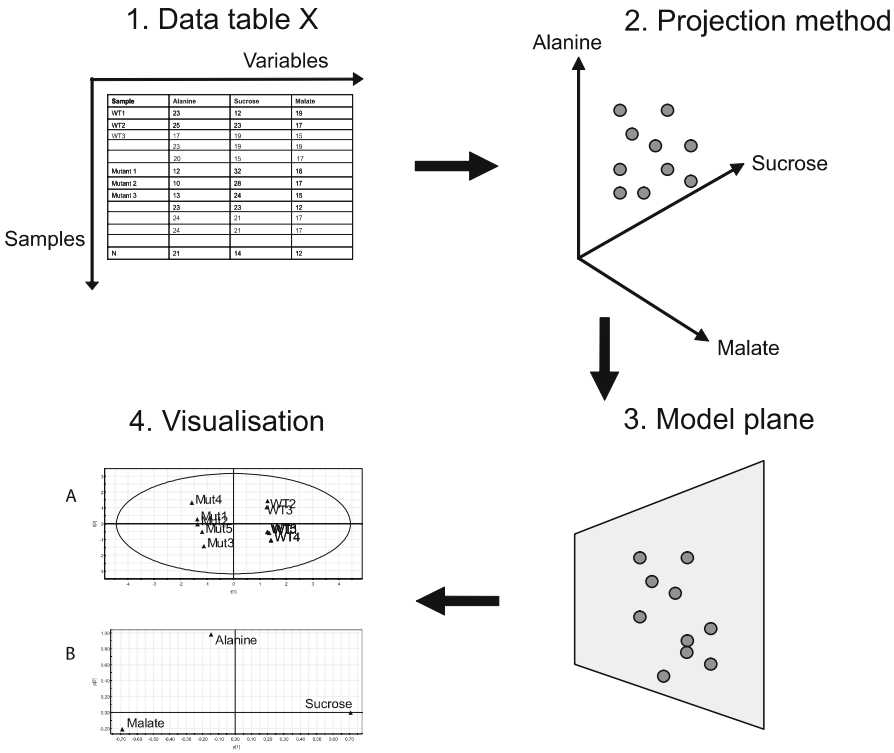


Fig. 3. (1) Each row (representing one biological sample) in a data table with K = 3 variables can be represented as one point in a K = 3 dimensional space. The position of that point is given by the coordinates given by the values in each of the K = 3 variables. (2) Repeating this for all rows (samples) in a data table produces a swarm of points in K = 3 dimensional space. Points (samples) that are close to each other have more similar biological properties than points that are far apart. (3) Projection methods such as PCA, finds a representative low-dimensional plane (here two-dimensional) that is a good summary of the variation in the X data table (swarm of points). (4) This model plane can then be visualised in scatter plots (A) and provides an overview, e. g. if there are any groupings, trends or outliers in the data. For example in the figure (A) there is a clear separation between the *Arabidopsis* wild type and mutant. It is also possible to understand the reason for this separation by looking at the direction of the model plane with respect to the original axes (original variables). These are summarized in the PCA model loadings, P (B)

loading plot (Fig. 3). This is a powerful tool for understanding the underlying patterns in the data.

The PCA model can be expressed as

$$\text{Model of X: } X = TP^T + E$$

where T are the scores, P defines the loadings, and E represent the residual matrix. The residual matrix E contains the residuals for each sample between

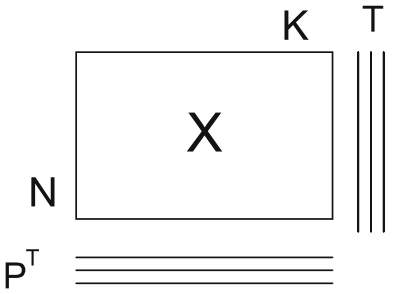


Fig. 4. PCA summarise all variation in X into a few new variables called scores T. These new variables are linearly weighted combinations of the original X-variables. The loadings P contain the weights used for each X-variable and thus reveal the influence of individual X-variables

its point in K-dimensional space and its point on the model plane. The residuals are important for detection of outliers and for defining the model boundaries (see Fig. 4).

2.5 Partial Least Squares Projections to Latent Structures (PLS)

The PLS method is used instead of the PCA method when additional knowledge about each sample exists, the Y matrix, e. g. genotype of each sample (wild type/mutant). The sample information according to the design matrix from the Design of Experiments (see Sect. 2.1) is often used as a Y matrix. Hence, PLS represents the regression analogy of PCA working with two matrices, X and Y (Wold et al. 1984). It is one of the most common methods when a quantitative relationship between a descriptor matrix X and a response matrix Y is sought. The Y matrix can contain both quantitative (e. g. glucose concentration) and qualitative (genotype) information. This additional sample information in Y is used by the PLS method to focus the model plane to capture the *Y-related variation* in X, e. g. separation between genotypes, rather than providing an overall view of *all variation* in the data as done by the PCA model. In addition, the PLS method can also be used to predict the properties (Y-values) of new unknown samples, e. g. predict the glucose concentration or genotype.

The Y matrix consists of the same number of rows as the X matrix. Each column in Y indicate a certain property, e. g. glucose concentration or genotype for each sample. When Y contains qualitative information such as genotype, the number of columns in Y equals the number of classes. Each row in Y describes the group membership for that sample where "1" indicates class belonging for that sample and "0" does not. When Y is qualitative, the PLS method is called PLS Discriminant Analysis (PLS-DA), to distinguish it from the situation when Y is quantitative.

3 Example: Metabolomics Study on Arabidopsis Mutants

We will work through a metabolomics example using GC/MS data from the analysis of *Arabidopsis* extracts. Shoots of higher plants are characterized by axillary branching, where the shoot branches develop from shoot meristems located between a leaf and the shoot stem. The control of axillary shoot growth (branching) is not well understood, but it is known that several internal factors such as the plant hormones IAA and cytokinins are involved (McSteen and Leyser 2005). Mutations screens in *Arabidopsis* have identified four loci involved in the repression of axillary bud growth, *MAX1–4*. Based on the mutants, it is now suggested that an unknown transmittable substance might be involved in controlling branching (see McSteen and Leyser 2005). The biosynthesis of this compound in *Arabidopsis* is catalyzed by a number of MAX (more-axillary growth) proteins.

We have used a metabolomics approach to classify and identify the metabolic differences between the MAX-mutants. Root samples from WT, max3 and max4 mutants were analysed by GC/TOFMS as described by Gullberg et al. (2004). The GC/MS data was processed by hierarchical multivariate curve resolution (Jonsson et al. 2005a), and the obtained X-matrix was thereafter subjected to PCA and PLS-DA analysis. The GC/MS processing resulted in 514 resolved peak areas. Log transformation, column centering and scaling to unit variance was done on the resolved peak areas (X-matrix) prior to modeling and two dummy Y-variables were constructed based on the class belonging of each sample to the

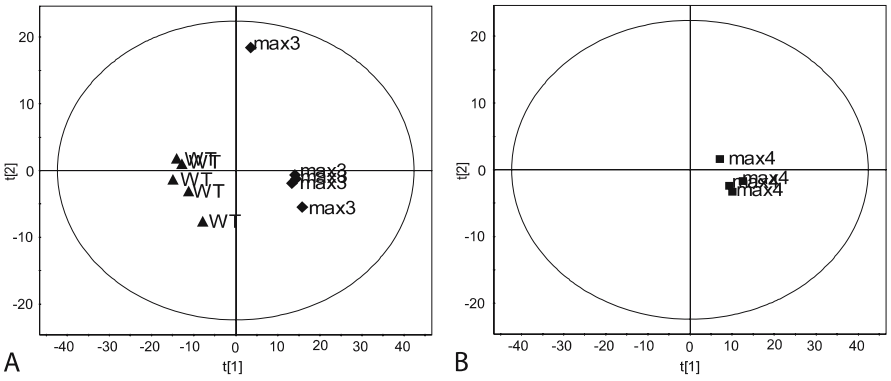


Fig. 5. A PLS-DA score-plot from the analysis of metabolite profiles in roots of *Arabidopsis* WT, max3 and max4. The PLS-DA model is based on WT and max3. The X-matrix was centered and scaled to unit variance. The explained variation in the X-matrix (R^2X) is 0.74, the explained variation in the Y matrix (R^2Y) is 0.99 and the predictive ability according to sevenfold cross-validation (Q2) is 0.84. R^2X is the cumulative modelled variation in X, R^2Y is the cumulative modelled variation in Y and Q^2Y is the cumulative predicted variation in Y, according to cross-validation. The range of these parameters is 0–1, where 1 indicates a perfect fit. **B** Based on the model max4 samples were predicted into the model showing that max3 and max4 are very similar regarding metabolic content (compare position score plot in **A**)

genotypes, WT and max3. The PLS-DA model score plot is shown in Fig. 5A. The score plot reveals the relationship among the samples. It is clear from the figure that the model plane displays a clear separation of the two genotypes.

To validate the model results, predictions were made for the genotype max4, using the calculated PLS-DA model based on the other sample-set (WT and max3). The results, shown in the obtained PLS-DA score plot (Fig. 5B) predicted that the max4 is closer to max3 than WT. This is consistent with the facts that max3 is very similar to the max4 genotype, where the MAX3 and MAX4 proteins use the same substrate (Schwarz et al. 2005). Interpretation of the first weight vector (w1) from the PLS-DA model, as described by Trygg and Wold (2002), together with the 99% confidence intervals calculated using jack-knifing (Martens and Martens 2000), highlighted 64 significant variables (metabolites) differing between WT and max4. The importance of these metabolites is a part of *biological validation* of the data set. The *statistical validation* was done by prediction of the max3 mutants into the WT/max4 model. Both type of validation is of importance for validating the multivariate data set.

4 Summary and Future Prospectives

Multivariate projection methods, e. g. PCA and PLS, represent a useful and versatile technology to modelling, monitoring and prediction of complex problems and data structures encountered within metabolomics and other 'omics' disciplines. The common denominator is that high complexity data tables are generated and that these data tables can be analysed and interpreted by means of chemometric methods. The principal component analysis (PCA) method summarizes the variation in a data table X into a model plane (the scores T). A scatter plot of these scores gives an overview of the samples (observations) and how they relate to each other, e. g. if there are groupings or trends or deviating samples and so on. In order to interpret the patterns found in a score plot one examines the corresponding loading plot (P). The loadings P reveal how each variable contributes to the separation among samples in the model plane and also gives insights into the relative importance of each variable.

However, one fundamental property is that the data does contain relevant information regarding our biological question. In other words, how to maximise the information content in the data? The traditional way to Change One Factor at a Time, i. e. the COST approach, is not recommended. Design of Experiments (DOE) is the methodology of how to conduct and plan experiments in order to maximize information in the data in the fewest number of runs. A proper experimental design will reveal the influence of each of the different factors separately and also any interactions between them. DOE is the only feasible approach to separate cause and effect from each other. Therefore is DOE in combination with chemometrical analysis a powerful way of planning, conducting and evaluating metabolomics experiments.

One common discussion point in the analysis of "omics" data is how to correlate several types of data, usually with different data structures. Systems biology seeks to integrate information from multiple parts of a biological system in a holistic attempt to understand the whole system. There are still many obstacles and hurdles to overcome in order to succeed. One of these relates to how the actual integration of the different types of data will be done. Hence, the advancement of systems biology depends heavily on the ability to integrate multiple profiling techniques (e. g. transcriptomics, proteomics, GC/MS, LC-NMR). The current multivariate statistical methods (e. g. the PLS method) lacks the proper model structure to describe these types of data structures, because they focus only on the *correlation pattern* among multiple data tables (e. g. X = microarrays vs Y = metabolomics data) and not on the *non-correlated variation* among these data tables which, in a biological sense, can be of equal interest. It has also been demonstrated that, because of this, the interpretation of these models are negatively affected (Trygg and Wold 2002), e. g. positive correlation patterns are interpreted as negligible or even flipped and become negative. This is a fundamental problem as we certainly cannot expect that all variation in transcript and metabolite levels co-vary. Fortunately, recent advances in chemometrics provide the ability to compare multiple data sets with each other. Novel extensions of the PLS method, called O-PLS (Trygg and Wold 2002) and O2-PLS (Trygg 2002) contain the model structure to support both these features. In addition, the O2-PLS method is bi-directional which means that the flow of information can go in both ways, from X (e. g. microarray) to Y (e. g. metabolomics) and vice versa. Hence, the O2-PLS methodology will be important in selecting what genes or metabolites are important to do further experimentation upon, e. g. understanding biomarker patterns and selecting genes for knockout studies. The O2-PLS methodology can also be extended to more than two data tables, hence it nicely fits into the framework of a combined profiling approach.

Acknowledgements. The Swedish Research Council, Wallenberg Consortium North (WCN), the Kempe foundation, EU strategic funding, Knut and Alice Wallenberg Foundation (JT) and Strategic Research Funding (SSF) are acknowledged for financial support. Professor Ottoline Leyser, York, UK, for allowing us to show data from the max-mutant project, and Dr. Miyako Kusano, RIKEN Plant Science Centre, Yokohama, Japan for the initial analysis of metabolites in the max-mutants.

References

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nature Biotechnol 21:692–696

Andreev VP, Rejtar T, Chen HS, Moskovets EV, Ivanov AR, Karger BL (2003) A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. Anal Chem 75:6314–6326

Dunn WB, Bailey NJC, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130:606–625

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi and megavariate data analysis. Umetrics (www.umetrics.com), ISBN 91–973730-1-X

Gullberg J, Jonsson P, Nordström A, Sjöström, M, Moritz T (2004) Optimisation of preparation of plant samples for metabolic profiling by GC-MS. Anal Biochem 331:283–295

Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids - potential for pattern recognition and automated identification of metabolic disorders. Rapid Commun Mass Spectrom 13:279–284

Idborg-Björkman H, Edlund, PO, Kvalheim OM, Schuppe-Koistinen I, Jacobsson SP (2003) Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. Anal Chem 75:4784–4792

Jackson JE (1991) A users guide to principal components. Wiley, New York

Jonsson P, Gullberg J, Nordström A, Kowalczyk M, Sjöström M, Moritz T (2004) A strategy for extracting information from large series of non-processed complex GC/MS data. Anal Chem 76:1738–1745

Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, Marklund S, Sjöström M, Antti H, Moritz T (2005a) Highthroughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. Anal Chem 77:5635–5642

Jonsson P, Bruce SJ, Moritz T, Trygg J, Sjostrom M, Plumb R, Granger J, Maibaum E, Nicholson JK, Holmes E, Antti H (2005b) Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. Analyst 130:701–707

Lundstedt T, Seifert E, Abramo L, Thelin B, Nyström A, Pettersen J, Bergman R (1998) Experimental design and optimization. Chem Intel Lab Systems 42:3–40

Martens H, Martens M (2000) Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual Pref 11:5–16

McSteen P, Leyser O (2005) Shoot branching. Annu Rev Plant Biol 56:353–374

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L et al (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579:1332–1337

Schwartz S, Qin XQ, Loewen MC (2005) The biochemical characterization of two Carotenoid cleavage enzymes from *Arabidopsis* indicates that a carotenoid-derived compound inhibits lateral branching. J Biol Chem 279:46940–46945

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62:817–836

Trygg J (2002) O2-PLS for qualitative and quantitative analysis in multivariate calibration. J Chemometr 16:283–293

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometrics 16:119–128

Wold S, Ruhe A, Wold H, Dunn WJ III (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Statist Comput 5:735–743

II.3 Map Editor for the Atomic Reconstruction of Metabolism (ARM)

M. Arita[1,2], Y. Fujiwara[1], and Y. Nakanishi[3]

1 Introduction

In the systems study of biological networks, computational analysis is expected to contribute in three phases by (1) *model selection*, the formal definition of each pathway's role in the manifestation of the biological aspect under analysis, (2) *model refinement*, the estimation of model parameters to refine constructed mathematical models, and (3) *simulation and feedback*, the computer simulation for the feedback of the predicted results for a better understanding of the target mechanism(s) (Hood 2003; Arita et al. 2005). Of these, the first phase is of prime importance because it determines the target of the analysis and its abstraction level. In other words, in the *model selection* phase, an appropriate model is searched and selected among all hypothetical candidates. The process of model selection is often performed intuitively by researchers. For example, glycolysis, the best-known pathway module in energy metabolism, contains a sequence of ten biochemical reactions from glucose to pyruvate (Berg et al. 2002). Since the pathway is linear, it can be summarized as if it were a single reaction:

$$\text{Glucose} + 2\text{NAD}^+ + 2\text{ADP} + 2\text{Pi} \rightarrow 2\text{Pyruvate} + 2\text{ATP} + 2\text{NADH} + 2\text{H}^+$$

The behavior of the pathway may be mathematically described either as a set of ten reactions, or as a single, abstract reaction. The abstract model is preferred in explaining net ATP generation, whereas the ten-reaction model is used for metabolic simulations. In choosing a model, we must be aware of the trade-off between model accuracy and its description length. In general, any model inevitably loses its fit to its corresponding natural mechanism as its description becomes simpler. In glycolysis, the nine intermediate molecules in the pathway can be eliminated to obtain the net ATP model in return for sacrificing biochemical details (i. e., the intermediates) of the pathway.

However, when glycolysis is placed in a global metabolic network, the description of the intermediate molecules is no less important than that of the gateway molecules such as glucose and pyruvate. The question arises as to

[1] Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa, 277–8561 Japan, e-mail: arita@k.u-tokyo.ac.jp
[2] Institute for Advanced Biosciences, Keio University, and Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency
[3] Intec Web and Genome Informatics Corporation

Biotechnology in Agriculture and Forestry, Vol. 57
Plant Metabolomics (ed. by K. Saito, R.A. Dixon, and L. Willmitzer)

which criteria should be evaluated and chosen for an appropriate abstraction of the given network. In biology, the focus has been directed at modularity in terms of function and structure.

The introduction of modularity, i. e. the encapsulation of network details by specifying the input/output, yields many advantages. First, it simplifies the description of the given network and facilitates its understanding. In metabolic networks, biologists have used intuitive, functional concepts such as 'amino-acid biosynthesis' or 'nitrogen assimilation' without formal, logical definitions. Second, the introduction of modularity simplifies the static verification of the network, a required step that precedes quantitative analyses such as simulation. The stoichiometric balance is one such property that can be verified statically.

In glycolysis, the structural and functional modularity is clear, mainly because the pathway is not branched; the pathway structure is linear and all carbon atoms in glucose are mapped to pyruvate and the function is the decomposition (lysis) of glucose. In general, modularity is less straightforward in branching metabolic pathways. Since molecular moieties are split into or merged with multiple molecules, it is not easy to trace carbon and other atomic elements, let alone delineate modularity. In fact, the determination of molecules to be regarded as intermediates is context-dependent: focusing on different atoms changes the pathways to be traced and therefore the resulting modular decomposition. For multiply branching pathways there is no universally effective, definitive decomposition.

Can the modularity of metabolic networks be detected computationally? Many automated methods exploit the stoichiometry of biochemical reactions (Mavrovouniotis 1992); this is one solution for the formal decomposition into metabolic modules. Currently, however, the most widely accepted method for finding modularity is through network topology; verification is by visualization with functional annotations (Ravasz et al. 2002; Ma et al. 2004). This is why pathway modules remain intuitive. Although electronic circuits can be verified using Boolean logic, no such formal system has been developed for biological systems. Indeed, many software programs for genomic and proteomic networks address only visualization, and their formal analysis remains to be solved.

The basic concept of our metabolic map editor, introduced in this chapter, combines a formal description of metabolism (structural conversions and their stoichiometric conditions) with intuitive network visualization. Many visualization tools and databases provide a static view of a given metabolic network (Mendes 2002; Kanehisa et al. 2004; Keseler et al. 2005). Because a metabolic network is well investigated and has a traditionally accepted layout for metabolites and enzymatic reactions, fully automated layout algorithms that ignore such standard layouts do little to further our understanding of its properties. Rather, an interactive drawing tool that can edit and modify standard metabolic networks is needed. Only with interactive software can biologists derive species-specific pathway images and visualize their intuitive ideas.

2 Definition of Metabolic Information

In this section, we introduce the concept of handling metabolic information at the atomic scale, and show how atomic representation can advance the formal understanding of metabolism.

2.1 Definition of Metabolic Pathways

Functions that should be computationally supported by a metabolic map editor include (1) searching and adding alternative or new metabolic pathways, (2) superimposing genomic, proteomic, and metabolomic data onto the network, and (3) rearranging the network topology to accommodate the metabolism of a particular species of interest. To fulfill these criteria, a formal definition of metabolic pathways is required. Previous computational studies tended to present a metabolic network as a graph where nodes and edges correspond to metabolites and their biochemical reactions, and the pathway as a sequence of graph edges (Ravasz et al. 2002; Ma et al. 2004). However, from a biochemical perspective, edge sequences (or graph paths) defined in this manner do not necessarily correspond to metabolic pathways. Since molecular structures are transformed in the course of each reaction, adjacent graph edges may not share the common structural moiety that corresponds to the metabolic (or atomic) flux (Fig. 1).

To resolve the conflict between graph paths and biochemical pathways, an additional constraint must be introduced (Arita 2003). In this chapter, *a metabolic pathway* (pathway for short) from metabolite *X* to metabolite *Y* is defined as a sequence of reactions through which at least one atom (carbon, nitrogen, or sulfur) in *X* reaches *Y*. A metabolite *Y* is called *reachable from X* if there is a pathway from *X* to *Y*. This is a rather strict constraint because the

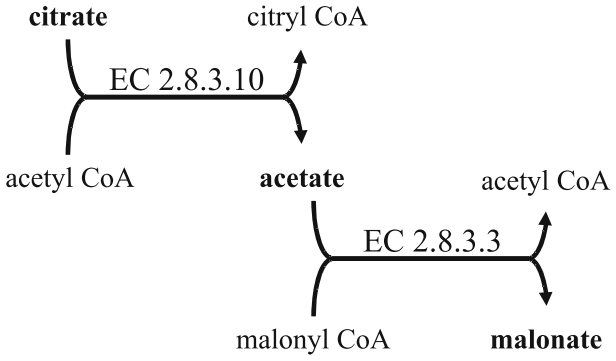


Fig. 1. Example of a biochemically inappropriate pathway from citrate to malonate. Analysis of molecular structures is required to enable the computer to detect that the citrate moiety is transferred to citryl CoA, rather than to acetate

conserved moiety throughout a pathway may not consist of carbon, nitrogen, or sulfur atoms; it may consist of oxygen or hydrogen atoms or even electrons. The map editor deals with only three types of elements because they can be computationally traced without ambiguity. The tracing of oxygen or hydrogen atoms is virtually impossible because the water molecule is involved in many reactions. The same is true for phosphates and metal ions that exist as free inorganics in a cell.

2.2 Representation of Pathways and Networks

In the atomic representation of a metabolism, each reaction is decomposed into a set of sub-structural correspondences called *atomic mappings* (Arita 2003). Each atomic mapping represents the transfer of a certain structural moiety in the course of biochemical reactions, and may be shared among multiple reactions. For example, atomic mapping between ATP and ADP, and between glutamate and α-keto-glutarate is prevalent in phosphate- and amino-transfer reactions, respectively (Fig. 2).

Conventionally, a metabolic pathway is thought of as a sequence of catalytic reactions classified by EC numbers. However, in the metabolic reconstruction from a whole genome sequence, a metabolic pathway is better viewed

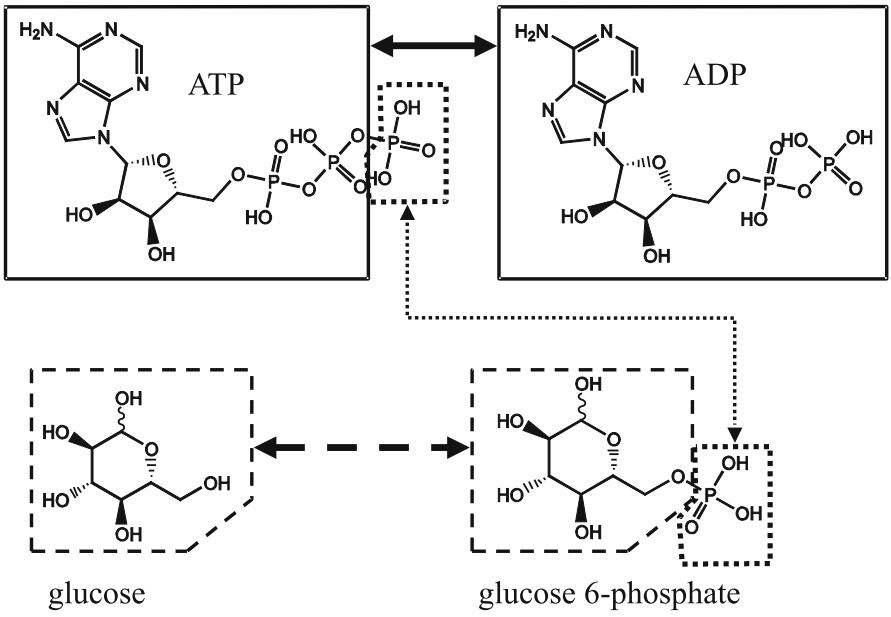


Fig. 2. Three atomic mappings in the reaction glucose + ATP = glucose 6-phospahte + ADP. The mapping between ATP and ADP (*shown in solid lines*) is common to all phosphorylation reactions with ATP and ADP

as a sequence of atomic mappings rather than as EC reactions. There are at least two reasons for this. First, the proposed atomic view can provide more candidate pathways in the reconstruction. Reconstruction based only on EC numbers often results in a set of incomplete pathways with multiple gaps. Such superficial gaps, however, may be filled with other reactions that share atomic mapping corresponding to the gap. Second, the proposed atomic view provides flexibility in choosing coenzymes. In theoretical analyses, the stoichiometric balance of reactions has been discussed as if coenzymes were fixed for all reactions (as in the traditional metabolic map) (Papin et al. 2004). In practice, however, their balance should be determined considering the net reaction balance in the network. For example, some NAD-dependent enzymes can also catalyze reactions using NADP, and the overall balance of their use depends on the total networking condition, not on individual reactions.

Thus, at least for the computational reconstruction of pathways, a metabolic network is better viewed as a set of atomic mappings rather than a set of EC-numbered reactions. In the biosynthesis of isoleucine and valine, for example, the same set of enzymes catalyzes 2-oxobutanoate and pyruvate to form isoleucine and valine, respectively. The only difference between the two pathways is a single alkyl group independent of the catalytic sequences in the biosynthesis. The decomposition into atomic mappings can explicitly describe such structural sharing between pathways.

3 Metabolic Map Editor

3.1 Overview

The design principle of the map editor is that users can flexibly integrate a sequence of atomic mappings (not reactions) into existing metabolic pathways to form metabolic maps (Fig. 3). First, users are expected to search metabolic pathways using the associated database that stores enzymatic reactions, their atomic mappings and molecular structures. The searched pathways (sequences of atomic mappings) are transferred to the main window where their layout can be freely edited as in a conventional graphical drawing editor. The advantage of our editor over conventional editors such as Microsoft PowerPoint is that users can import metabolic objects (e. g. compound structures and reactions) from the background database: although on-screen it appears as if only graphical objects for compound structures and reactions are imported, more information is processed in the background. For example, importing one enzymatic reaction on the screen implicitly invokes the integration of its associated atomic mappings into the already drawn metabolic map so that the route of any atom in the new reaction can be traced seamlessly on the resulting metabolic map.

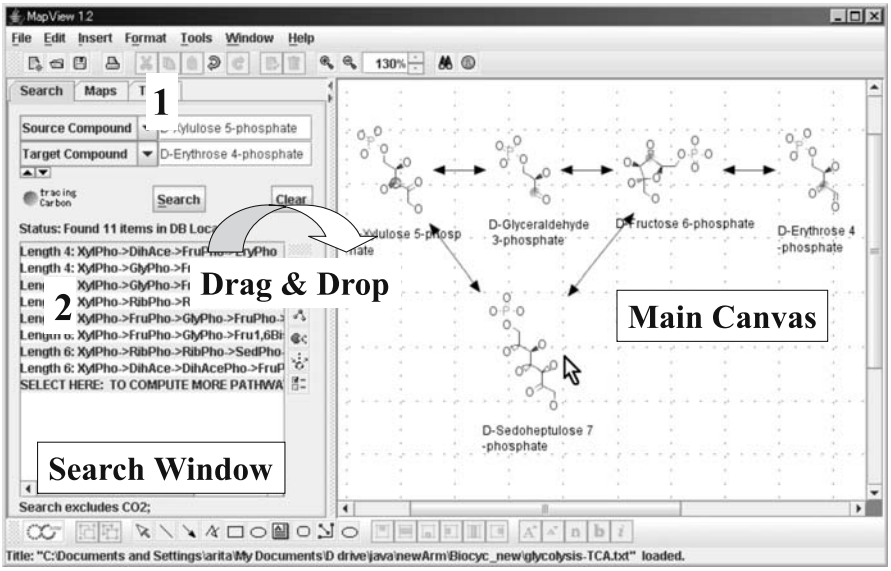


Fig. 3. Screen-shot of the Map Editor. Pathways are searched by typing molecular names in the input fields (*Number 1*). The search results are listed (*Number 2*). Users can drag pathways into the main canvas

3.2 Metabolic Database

The background database for atomic information stores three types of metabolic data: molecular structures, reaction formulas, and their atomic mappings. All data are freely available in text format from the website http://www.metabolome.jp/download.html.

Molecular structures are registered in the MOL-file format (MDL Information Systems; its description is downloadable from http://www.mdli.com/). The MOL-file format is the de facto standard to describe molecular structures; an example is shown in Fig. 4. Each MOL-file describes one molecular structure as a list of atoms with their XYZ coordinates and their chemical bondings. The chirality of carbon is specified using one integer value for each corresponding carbon atom. Information on the display of chirality (in thick and shaded lines) is specified using other integer values. The metabolic editor does not use the XYZ coordinates written in the format; rather, it applies the original drawing algorithm to assign XYZ positions (Arita 2005).

As in other metabolic databases, enzymatic reactions are described using compound names. Reaction formulas were obtained from the Enzyme Nomenclature of the International Union of Biochemistry and Molecular Biology (http://www.chem.qmw.ac.uk/iubmb/enzyme/). In each reaction, the order of molecules on the left- and right-hand side was manually rearranged so that the atomic mappings can be computationally detected by comparing molecu-

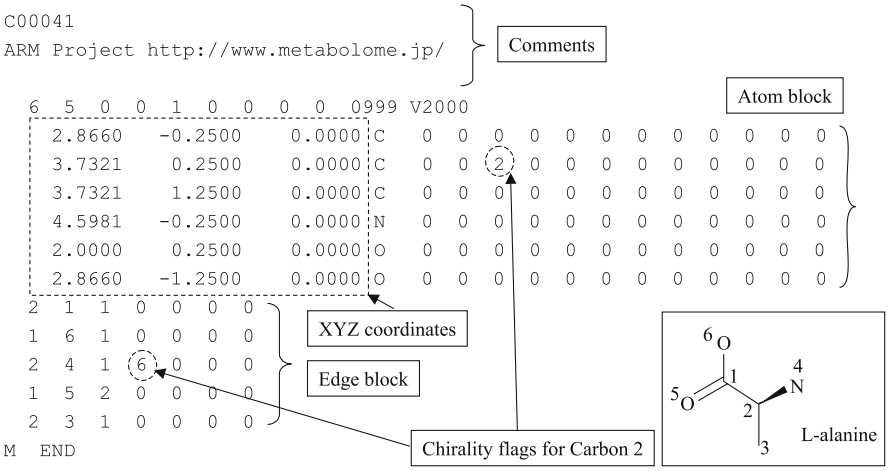


Fig. 4. MOL-file format for L-alanine. In the ARM database, carbon atoms in the atom-block are ordered according to the IUPAC positions for molecular structures. Only carbon atoms are correctly ordered

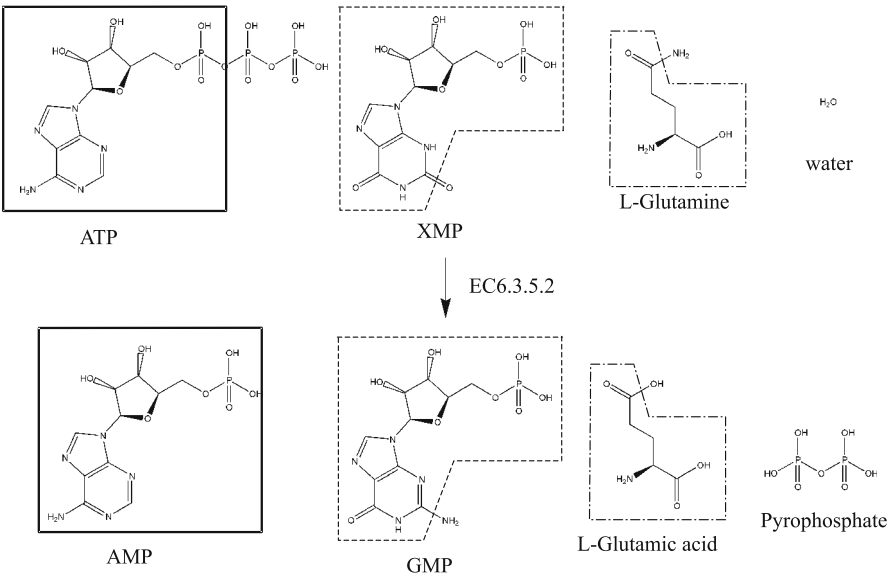


Fig. 5. Schematic view of reaction EC 6.3.5.2. The reaction formula is written as "ATP + XMP + L-glutamine + H_2O =>AMP + GMP + L-glutamic acid + pyrophosphate" in the database so that the molecular structures roughly correspond one-to-one (*top to bottom*)

lar structures sequentially from left to right (Fig. 5). For details on structure comparison to compute atomic mappings refer to Arita (2003).

Using molecular structures, atomic mappings were pre-computed for all registered reactions and, after manual verification, correct mapping results

were stored in the database. Details, including the accuracy of the mapping computation, were described previously (Arita 2003). Although the mapping was computed for all atomic elements except hydrogen, the results were registered only for carbon, nitrogen, and sulfur atoms due to ambiguities in the mapping of the rest of the elements.

3.3 Drawing Maps from Pathways

The map editor is equipped with a search engine for metabolic pathways. Given a source and target metabolites, the engine computes logically possible pathways between these metabolites from the shortest- to pathways of any length. Although pathway length is measured by the number of reaction steps, an arbitrary value can be assigned in the algorithm used. In other words, the engine can compute any pathway throughout which at least one carbon (or nitrogen, sulfur) atom is conserved. An arbitrary combination of pathways can be visualized by dragging a searched pathway into the main canvas window (Fig. 3). When a pathway is dragged into the canvas, it is merged with the already drawn network (Fig. 6). Although its initial layout is automatically assigned, a user can freely rearrange the orientation or location of any metabolic object by using the mouse.

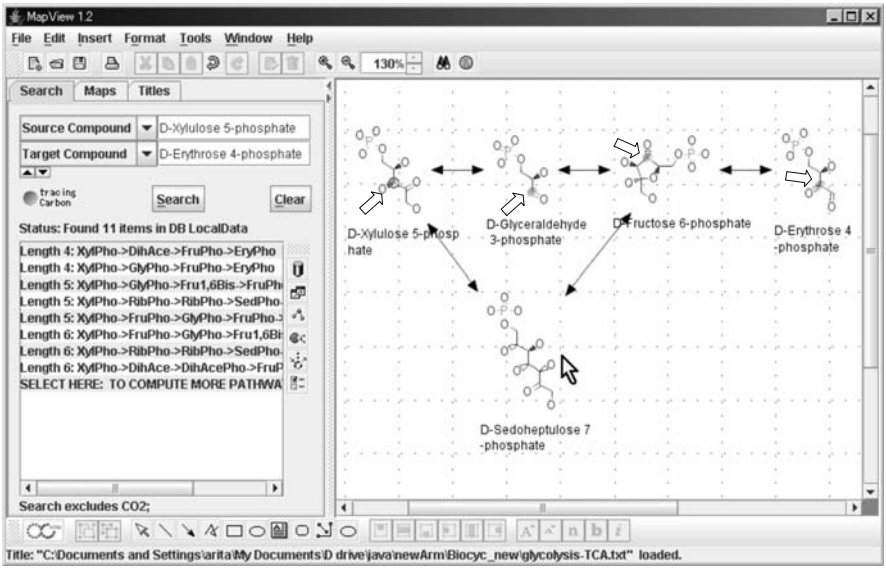


Fig. 6. The network generated by merging two pathways from xylulose 5-phosphate (X5P) to erythrose 4-phosphate. Every time a pathway is dropped, only the difference from the existing network is drawn. Carbon 3 in X5P is traced in this example (*shown with blank arrows*)

The unique function of the map editor is its ability to trace a particular atom on the map. Since each metabolic object on the map is linked with its atomic information in the database, the logical tracing of each atomic position is possible by transitive calculation of the atomic mappings in the network. A user needs only to mouse-click a particular atom on the network to see its traces (Fig. 6).

4 Applications

4.1 Carbon Flow in Cyclic Pathways

Metabolic pathways contain two types of cycles in terms of tracing atoms: cycles where carbon atoms are exchanged in each round (e. g. the tricarboxylic acid (TCA) cycle), and cycles where all carbon atoms are conserved (e. g. the urea cycle). If reversible, a single reaction catalyzing two identical molecules (e. g. 2 pyruvate = 2-acetolactate +CO_2) can form the former type of cycle by itself. Likewise, the latter type of cycle can be formed by any reversible reaction. Cyclic pathways may be biochemically meaningful, but in practice, their existence is problematic in searching metabolic pathways. Since pathways are searched and output according to the number of reaction steps in our system, short cycles drastically increase the number of spurious pathways with local loops. To eliminate such futile pathways, our pathway-search algorithm eliminates all pathways that visit the same molecule multiple times. However, this constraint is too strict for searching all possibly existing pathways. For example, users studying the TCA cycle may want to analyze carbon traces that go round the cycle multiple times.

To support the atomic analysis of cyclic pathways, a metabolic map editor is indispensable. First, users search the pathways of interest and paste them into the main window using the edit function (i. e., model selection). Then, a particular atom can be interactively traced within the selected set of reactions. Since the target model is highly constrained, it is feasible to compute pathways visiting the same compound multiple times.

4.2 Carbon Flow in Energy Metabolism

Because of the shared metabolites between the glycolytic and pentose phosphate pathways, the atomic traces of a particular carbon atom often become hard to follow. This is the case for the correspondence between the C-1 of glucose and the C-1 of pentose 5-phosphate. By clicking the corresponding atomic position in the map editor, the atomic trace can be grasped at a glance. Although the entry point of the pentose phosphate pathway decarboxylates the C-1 position of glucose, this position is identical to the C-1 of fructose 6-phosphate through glycolysis. Thus, in the pentose phosphate pathway, the C-1

of fructose 6-phosphate corresponds to the C-1 of sedoheptulose 7-phosphate, of xylulose 5-phosphate, and of ribose- and ribulose 5-phosphate. In fact, the C-1 of glucose corresponds to both the C-1 and C-6 positions of fructose 6-phosphate, and therefore to all C-1 and C-5 positions of pentose 5-phosphate. These positions are invariable throughout the glycolytic and pentose phosphate pathways.

4.3 Visualization of Lipid Metabolism

Recently, metabolome analysis has been facilitated due to the rapid technical progress made in mass spectrometry (MS). To detect lipid molecules, for example, an effective strategy is to couple MS with liquid chromatography-electrospray ionization (Houjou et al. 2004). More than 1000 glycerophospholipid species can be quantitated in a single assay in less than 2 h (R. Taguchi, personal communication). However, the efficient analysis of such large-scale data sets poses a vexing problem. Network visualization remains the first step for gaining an overview of the data; however, the traditional metabolic map is not suitable for visualization because it contains abstract notations. For example, the 'phosphatidyl group' contains two fatty acids of variable lengths (usually 12∼24 carbon atoms) and degrees of unsaturation (usually 0∼6 double bonds, depending on the length). Since experimentally confirmed fatty acids in a phosphatidyl group are comprised of more than 30 species, the number of actual phosphatidyl species may be as many as its square, i. e., $\approx$ 1000. To visualize the distribution of the spectrum of molecular species, our map editor supports an interactive instantiation of abstract moieties. For each abstract notation using 'R-group', a user can assign a list of molecules as its possible instantiation. The map editor can also display an integer value for each molecule (such as the concentration, mass, logP, etc.). Given the list of possible instantiations for each R-group and their corresponding concentrations (i. e., metabolomic data), the editor can display the percentage fraction of candidate molecules. When multiple R-groups exist, the amount to be displayed will be the integration of all possible assignments. In a phosphatidyl group, various fatty acids can be linked at R1 and R2 positions of glycerol phosphate. With a mouse click, the candidate list for the R1 (or R2) position is displayed together with the relevant percentage fractions. The percentage for a docosahexanoic acid (DHA) in the R1 group, for example, is calculated as the sum of all phosphatidyl molecules that have DHA at R1. When DHA is chosen for R1, the percentage list for R2 consists of fully instantiated molecular species that have DHA at R1 and another fatty acid at R2.

Due to the abstract notations for molecules, lipid metabolism is a particularly unspecific part in the traditional metabolic map. The computer-assisted metabolic map is indispensable to visualize the metabolomic data of such pathways.

5 Conclusions

The map editor is not only a tool for visualizing metabolic pathways, but is a necessary component for the systematic and modular understanding of species- and context-dependent metabolic networks. Since the software system is linked with atomic-level information in the background database, users can trace any atomic position on any metabolic network they draw. It is a desirable realization of a pathway database. Most web-based pathway databases do not support the users' own arrangement of networks, although in computer science, the definition of a database system is ' a collection of information organized in such a way that a computer program can quickly select desired data in a desired arrangement'. Our map editor compensates for this drawback, and represents a step forward to a more flexible analysis of large-scale biological information.

Acknowledgements. The ongoing analysis of lipid metabolism is a joint effort with Prof. Ryo Taguchi at The University of Tokyo. The authors thank Ursula Petralia for editing the manuscript. This work was supported by The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research on Priority Areas.

References

Arita M (2003) *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. Genome Res 13(11):2455–2466

Arita M (2005) Introduction to the ARM database: database on chemical transformations in metabolism for tracing pathways. In: Tomita M, Nishioka T (eds) Metabolomics: the frontier of systems biology. Springer, Berlin Heidelberg New York, pp 193–211

Arita M, Robert M, Tomita M (2005) All systems go: launching cell simulation fueled by integrated experimental biology data. Curr Opin Biotechnol 16(3):344–349

Berg JM, Tymoczka JL, Stryer L (2002) Biochemistry, 5th edn. Freeman, New York

Hood L (2003) Systems biology: integrating technology, biology, and computation. Mech Ageing Dev 124:9–16

Houjou T, Yamatani K, Nakanishi H, Imagawa M, Shimizu T, Taguchi R (2004) Rapid and selective identification of molecular species in phosphatidylcholine and sphingomyelin by conditional neutral loss scanning and MS3. Rapid Commun Mass Spectrom 18(24):3123–3130

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32 (Database Issue):D277–D280

Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33 (Database Issue):D334–337

Ma HW, Zhao XM, Yuan YJ, Zeng AP (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. Bioinformatics 20(12):1870–1876

Mavrovouniotis ML (1992) Computer-aided synthesis of biochemical pathways. Biotechnol Bioeng 36:1119–1132

Mendes P (2002) Emerging bioinformatics for the metabolome. Brief Bioinform 3(2):134–145

Papin JA, Reed JL, Palsson BO (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem Sci 29(12):641–647

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–1555

Subject Index

Biotechnology in Agriculture and Forestry

Volumes already published

Volume 1: Trees I (1986)
Volume 2: Crops I (1986)
Volume 3: Potato (1987)
Volume 4: Medicinal and Aromatic Plants I (1988)
Volume 5: Trees II (1989)
Volume 6: Crops II (1988)
Volume 7: Medicinal and Aromatic Plants II (1989)
Volume 8: Plant Protoplasts and Genetic Engineering I (1989)
Volume 9: Plant Protoplasts and Genetic Engineering II (1989)
Volume 10: Legumes and Oilseed Crops I (1990)
Volume 11: Somaclonal Variation in Crop Improvement I (1990)
Volume 12: Haploids in Crop Improvement I (1990)
Volume 13: Wheat (1990)
Volume 14: Rice (1991)
Volume 15: Medicinal and Aromatic Plants III (1991)
Volume 16: Trees III (1991)
Volume 17: High-Tech and Micropropagation I (1991)
Volume 18: High-Tech and Micropropagation II (1992)
Volume 19: High-Tech and Micropropagation III (1992)
Volume 20: High-Tech and Micropropagation IV (1992)
Volume 21: Medicinal and Aromatic Plants IV (1993)
Volume 22: Plant Protoplasts and Genetic Engineering III (1993)
Volume 23: Plant Protoplasts and Genetic Engineering IV (1993)
Volume 24: Medicinal and Aromatic Plants V (1993)
Volume 25: Maize (1994)
Volume 26: Medicinal and Aromatic Plants VI (1994)
Volume 27: Somatic Hybridization in Crop Improvement I (1994)
Volume 28: Medicinal and Aromatic Plants VII (1994)
Volume 29: Plant Protoplasts and Genetic Engineering V (1994)
Volume 30: Somatic Embryogenesis and Synthetic Seed I (1995)
Volume 31: Somatic Embryogenesis and Synthetic Seed II (1995)
Volume 32: Cryopreservation of Plant Germplasm I (1995)
Volume 33: Medicinal and Aromatic Plants VIII (1995)
Volume 34: Plant Protoplasts and Genetic Engineering VI (1995)
Volume 35: Trees IV (1996)
Volume 36: Somaclonal Variation in Crop Improvement II (1996)
Volume 37: Medicinal and Aromatic Plants IX (1996)
Volume 38: Plant Protoplasts and Genetic Engineering VII (1996)
Volume 39: High-Tech and Microprogation V (1997)
Volume 40: High-Tech and Microprogation VI (1997)
Volume 41: Medicinal and Aromatic Plants X (1998)
Volume 42: Cotton (1998)
Volume 43: Medicinal and Aromatic Plants XI (1999)
Volume 44: Transgenic Trees (1999)
Volume 45: Transgenic Medicinal Plants (1999)
Volume 46: Transgenic Crops 1I (1999)
Volume 47: Transgenic Crops II (2001)
Volume 48: Transgenic Crops III (2001)

Volumes 1-48 were edited by Y.P.S. Bajaj†

Volume 49: Somatic Hybridization in Crop Improvement II (2001)
T. Nagata and Y.P.S. Bajaj (Eds.)
Volume 50: Cryopreservation of Plant Germplasm II (2002)
L.E. Towill and Y.P.S. Bajaj (Eds.)
Volume 51: Medicinal and Aromatic Plants XII (2002)
T. Nagata and Y. Ebizuka (Eds.)
Volume 52: Brassicas and Legumes: From Genome Structure to Breeding (2003)
T. Nagata and S. Tabata (Eds.)
Volume 53: Tobacco BY-2 Cells (2004)
T. Nagata, S. Hasezawa, and D. Inzé (Eds.)
Volume 54: *Brassica* (2004)
E.C. Pua and C.J. Douglas (Eds.)
Volume 55: Molecular Marker Systems in Plant Breeding and Crop Improvement (2005)
H. Lörz and G. Wenzel (Eds.)
Volume 56: Haploids in Crop Improvement II (2005)
C.E. Palmer, W.A. Keller, and K.J. Kasha (Eds.)
Volume 57: Plant Metabolomics (2006)
K. Saito, R.A. Dixon, and L. Willmitzer (Eds.)

Volumes in preparation

Tobacco BY-2 Cells: A New Treatise
T. Nagata, K. Matsuoka, and D. Inzé (Eds.)

Transgenic Crops IV
E.C. Pua and M.R. Davey (Eds.)

Transgenic Crops V
E.C. Pua and M.R. Davey (Eds.)

Transgenic Crops VI
E.C. Pua and M.R. Davey (Eds.)